Reprojection of the Conjugate Directions in the ABS Class Part I

József Abaffy

Óbuda University, Institute of Applied Mathematics
H-1034 Budapest, Bécsi út 96/b, Hungary
abaffy.jozsef@nik.uni-obuda.hu

Abstract. In the paper we introduce a modification of the Kahan-Parlett "twice is enough" [20] algorithm for conjugate direction algorithms. We apply the developed algorithm for the computation of conjugate directions in three subclasses of the ABS methods. In this part of the paper we give theoretical results and a preliminary test result as well. In the second part of our paper we test some elements of Subclass S2, while in the third part Subclass S6 and S7 will be examined. In this last part we give the conclusions regarding the reprojection of conjugate directions in the ABS classes

Keywords: twice is enough, reprojection of conjugate direction methods, ABS methods

1 Introduction

A very important question in Linear Algebra is the error propagation of algorithms and monitoring the error during the execution. There are many works in this field. We refer to [3], [5], [7], [8], [9], [10], [11], [13], [14], [16], [20], [22], [23] , [24] and others.

The reprojection technic can be applied in every class of the ABS methods. This reprojection improves accuracy in general, but the rounding errors could vanish these improvements. Therefore, it is very important to set up conditions for the reprojection. For orthogonal vectors the "twice is enough" method was developed by Parlett and Kahan [20]. The reprojection technic cannot be applied trivially to the conjugate direction methods like Lánczos [18], [19], Hestenes Stiefel [15] and others.

In this paper we develop a reprojection algorithm to conjugate directions and we give the necessary theoretical results to apply it to different subclasses (S2 and S6) of the ABS class.

We consider the classical conjugate gradient problem. A is supposed to be symmetric positive definite matrix throughout this paper.

We have to mention that we do not consider A^TA or AA^T conjugate direction methods, such as the CG algorithm applied to normal equations, Craig's method, CR, GCR, ORTHOMIN, ORTHODIR, GMRES methods and so on. For these methods see, for example [12], [1]. These problems and algorithms will be considered in a following paper.

2 Theoretical background

In this section we give the conjugate reprojection version of the Parlett and Kahan (PK) [20] method using the ABS class [1].

2.1 Parlett-Kahan type reprojection of conjugate directions in the ABS class

First we present the scaled ABS class which will be used later frequently.

Let us consider the following scaled system

$V^TAx = V^Tb$

where $A \in \Re^{m,n}$, $V \in \Re^{m,m}$ is an arbitrary non-singular matrix, $b \in \Re^m$ and $x \in \Re^n$.

The class of the scaled ABS algorithm

Algorithm 1. *Step 1 Set* $x_1 \in \Re^n$, $H_1 = I \in \Re^{n,n}$ *where I is the unit matrix,* $i = 1$, *and* $iflag = 0$.

Step 2 Let $v_i \in \Re^n$ *be arbitrary save that* $v_1, ..., v_i$ *be linearly independent. Compute the residual error vector* $r_i = Ax - b$. *If* $r_i = 0$, *stop* x_i *solves the system. Otherwise, compute the scalar* $\tau_i = v_i^T r_i$ *and the vector* $s_i = H_iA^Tv_i$.

Step 3 If $s_i \neq 0$, *go to Step 4; if* $s_i = 0$ *and* $\tau_i = 0$, *set* $x_{i+1} = x_i$, $H_{i+1} = H_i$, $iflag = iflag + 1$, *and if* $i < m$, *go to Step 6; otherwise, stop; if* $s_i = 0$ *and* $\tau_i \neq 0$, *set* $iflag = -i$ *and stop.*

Step 4 Compute the search direction p_i *by*

$p_i = H_i^Tz_i$

where $z_i \in \Re^n$ *is arbitrary saving for* $z_i^TH_iA^Tv_i \neq 0$.

Step 5 Update the approximate of the solution by

$x_{i+1} = x_i - \alpha_i p_i$

where the step size α_i *is given by*

$$\alpha_i = \frac{\tau_i}{v_i^T A p_i}$$

if $i = m$*, stop and* x_{i+1} *is the solution of the equations.*

Step 6 Update the matrix H_i *by*

$$H_{i+1} = H_i - \frac{H_i A^T v_i * w_i^T H_i}{w_i^T H_i A^T v_i} \tag{1}$$

where w_i *is arbitrary but the denominator must be non-zero.*

Step 7 Set $i = i+1$ *and go to Step 2.*

The properties of this algorithm can be found in [1]. Further we do not use the index i only if it is necessary.

We shall use two formulas of the H^T projection matrix. One of them can be obtained from (1) and the other is

$$\overline{H}^T = I - W * Q^{-T} * V^T * A \tag{2}$$

where W and V contain $w_1, ..., w_i$ and projection vectors $v_1, ..., v_i$ respectively computed and $Q^{-T} = (W^T A^T V)^{-T}$ until the actual step. For (2) see formula (7.20) or (7.59) of [1].

Now we are able to present the ABS PK type conjugate direction method. First we define the error vectors. Let the error vector $e' = x' - p$ satisfy $e'^T A e' \leq \varepsilon z^T A z$ where x' is the approximation of $p = \overline{H}^T z$ and $e" = x" - p$ satisfy $e"^T A e" \leq \varepsilon z^T A z$ where $x"$ is the approximation of $p = \overline{H}^T x'$ and ε is some tiny positive ε independent of z and A.

Let κ be any fixed value in the range $[1/(0.83-\varepsilon), 0.83/\varepsilon]$. Using the notation of the algorithm "twice is enough" we give.

Algorithm 2. *ABS Conjugate Direction of Parlett Kahan type (ABS_CD_PK)*

Case 1. If $x'^T A x' > z^T A z/\kappa$ *accept* $x = x'$ *and* $e = e'$*, otherwise compute* $x" = \overline{H}^T x'$ *to get* $x"$ *with error* $e"$ *and go to Case 2.*

Case 2. If $x"^T A x" \geq x'^T A x'/\kappa$ *accept* $x = x"$ *and* $e = e"$.

Case 3. If $x"^T A x" < x'^T A x'/\kappa$ *accept* $x = 0$ *and* $e = -p$.

As in the different subclasses the projection matrix H is calculated with different formulas and the theorems which ensure the accuracy of the ABS_CD_PK algorithm will be given there.

2.2 The class of the conjugate direction ABS algorithm (S2)

In this section, we study the S2 subclass of scaled ABS algorithm. Instead of the original equation $Ax = b$, where $A \in \Re^{m,n}$, $b \in \Re^m$, $x \in \Re^n$ consider the scaled equations

$$V^T Ax = V^T b \tag{3}$$

where $V = (v_1, \cdots, v_m) \in \Re^{m,n}$ is a non-singular matrix. The subclass S2 generates conjugate directions is defined by the formula

$v_i = p_i$

Note that we still have two arbitrary vectors z_i and w_i.

We recall Theorem 8.6 of [1] which state that the S2 subclass generates conjugate directions.

Theorem 1. *Let A be symmetric and positive definite. Then the subclass S2 where $v_i = p_i$ generates A conjugate search vectors and the iterate x_{i+1} minimizes over the linear variety $x_1 + Span(p_1, ..., p_i)$ the convex quadratic function*

$F(x) = (x - x^*)^T A(x - x^*)$

where x^* is the unique minimum point of $F(x)$.

Note that it is a special case of Theorem 7.17 in [1].

Now we prove a theorem which shows the effect of the reprojection with ABS_CD_PK.

Theorem 2. *The vector x computed by the ABS_CD_PK algorithm ensures that*

$e^T Ae^T \leq \varepsilon z^T Az^T + O(\varepsilon^2)$

and

$\left|p_0^T Ax\right| \leq \kappa\varepsilon p_0^T Ap_0 x^T Ax + O(\varepsilon^2)$.

Proof. We present those steps of the proof only which use the H projection matrix. The other parts of the proof are the same as in [20].

Case 1.

$e^T Ae = e'^T Ae' \leq \varepsilon z^T Az$

$\left|p_0^T Ax\right| = \left|p_0^T Ax'\right| = \left|p_0^T A(e' - p)\right| = \left|p_0^T Ae' - p_0 Ap\right| = \left|p_0^T Ae'\right|$ because of the conjugacy the second term is zero. Now, by applying the Cauchy–Schwartz inequality we get $\left|p_0^T Ae'\right| \leq \left\|p_0^T A^{1/2}\right\| \left\|A^{1/2} e'\right\| = \left(p_0^T A^{1/2} A^{1/2} p_0\right) \left(e'^T A^{1/2} A^{1/2} e'\right)$

$= \left(p_0^T Ap_0\right) \left(e'^T Ae'\right) \leq \left(p_0^T Ap_0\right) \varepsilon \left(z^T Az\right) =$

$\left(p_0^T Ap_0\right) \varepsilon\kappa \left(x'^T Ax'\right) = \varepsilon\kappa \left(p_0^T Ap_0\right) \left(x^T Ax\right)$ because of the true branch of Case 1.

Case 2.

$\left|p_0^T Ax\right| = \left|p_0^T Ax"\right| = \left|p_0^T A(e"+p)\right| = \left|p_0^T Ae"+p_0^T Ap\right| = \left|p_0^T Ae"\right|$

as the second term is zero because of the conjugacy

$= \left(p_0^T A^{1/2} A^{1/2} p_0\right)\left(e"^T A^{1/2} A^{1/2} e"\right) \leq \left(p_0^T Ap_0\right)\varepsilon\left(x'Ax'\right)$

and again from the true branch we get

$\leq \left(p_0^T Ap_0\right)\varepsilon\left(x"Ax"\right) \leq \kappa\varepsilon\left(p_0^T Ap_0\right)(xAx)$

On the other hand

$$\left(e^T Ae\right) = (x"-p)^T A\,(x"-p) \tag{4}$$

where $p = \left(I - W * Q^{-T} * P^T * A\right)x'$ therefore

$x"-p = e"+p+\left(I-W*Q^{-T}*P^T*A\right)x'-p$

$= e"+\left(I-W*Q^{-T}*P^T*A\right)(e'+p)-p$ and because of the conjugacy

$= e"+p+\left(I-W*Q^{-T}*P^T*A\right)e'-p = e"+\overline{H}e'$. Substituting it in (4) we get

$\left(e^T Ae\right) = \left(e"+\overline{H}e'\right)^T A\left(e"+\overline{H}e'\right)$

$= e"^T Ae"+e'^T\overline{H}Ae"+e"A\overline{H}e'+e'^T\overline{H}A\overline{H}e'$

$\leq \frac{\varepsilon}{\kappa}z^T Az+\|e'\|\left\|\overline{H}\right\|\|A\|\|e"\|+\|e"\|\|A\|\left\|\overline{H}\right\|\|e'\|+$

$\|e'\|\left\|A\overline{H}\right\|\|A\|\left\|\overline{H}H\right\|\|e'\|.$

Suppose that $\left\|\overline{H}\right\| \leq K$ then

$\leq \frac{\varepsilon}{\kappa}z^T Az+K\|A\|\|e'\|\|e"\|+K\|A\|\|e'\|\|e"\|+K^2\|A\|^2\|e'\|^2 \leq$

$\frac{\varepsilon}{\kappa}z^T Az+2K\|A\|\varepsilon z^T Az*\varepsilon x'^T Ax'+K^2\|A\|^2\varepsilon^2(z^T Az)^2.$

As now $x'^T Ax' \leq \frac{1}{\kappa}z^T Az$ we can continue

$\leq \frac{\varepsilon}{\kappa}z^T Az+2K\|A\|\frac{\varepsilon^2}{\kappa}\left(z^T Az\right)^2+\varepsilon^2K^2\|A\|^2\left(z^T Az\right)^2 =$

$\frac{\varepsilon}{\kappa}z^T Az+2K\|A\|\varepsilon^2\left(z^T Az\right)^2\left(\frac{1}{\kappa}+K\|A\|\right) = \frac{\varepsilon}{\kappa}z^T Az+O(\varepsilon^2)$

$\leq \varepsilon z^T Az+O(\varepsilon^2)$

as $\kappa > 1$ will be suggested to use.

Case 3. As $\left|p_0^T Ax\right| = 0$, it is enough to prove that $b^T p = b^T a = 0$, where $a = (I-W*Q^{-T}*P^T*A)e'$ and $b^T = e'^T(W*Q^{-T}*V^T*A)$. Indeed,

$b^T p = e'^T(W*Q^{-T}*V^T*A)*\left(I-W*Q^{-T}*P^T*A\right)x'$

$= e'^T(W*Q^{-T}*V^T*A-W*Q^{-T}*V^T*A*W*Q^{-T}*P^T*A)x' = 0.$

The proof for the case $b^T a = 0$ is similar to $b^T p = 0$. □

Note that the term which contains ε^2 can influence the estimation if $\|A\|$ is big. This phenomena will be observed during the tests of different algorithms.

Consider now the symmetric matrix projection case.

Symmetric matrices H_i are obtained for example with $H_1 = I$ where I is the unit matrix and w_i given by

$$w_i = \frac{Ap_i}{\|H_i A p_i\|_2^2} \tag{5}$$

In this case (5) is well defined.

Theorem 3. *If* $q_i = \frac{H_i A^T p_i}{\|H_i A p_i\|_2}$. *Then* $q_i^T q_j = 0$ *for* $i, j = 1, ..., n$

Proof. Let $j < i$ be. Then

$$q_i^T q_j = \frac{p_i^T H_i^T H_j A^T p_j}{\|H_i A p_i\|_2^2} = p_i^T H_i^T A^T p_j \, \|H_i A p_i\|_2^2 = \frac{p_i^T H_i A^T p_j}{\|H_i A p_i\|_2^2} = 0$$

because H_i is symmetric matrix $Null(H_i) = \left\{A^T p_1, ..., A^T p_{i-1}\right\}$ and the denominator is different from zero. The same argument is valid for the case $j > i$. □

Let $Q_i = [q_1, ..., q_i]$ be then we can obtained a block form of the projection matrix H_{i+1}

$$H_{i+1} = H_1 - Q_i Q_i^T. \tag{6}$$

It is important to note that the conjugate directions $p_i, i = 1, ..., n$ are generated by orthogonal column vectors of Q_i. Now we can only choose vectors z_i arbitrary. As the matrix update (8.24) of [1] takes an important role in some algorithms we present it now:

$$H_{i+1} = H_i - \frac{H_i A^T p_i p_i^T}{p^T A p_i} \tag{7}$$

where we used the idempotency of H_i. We present the chosen cases both for the symmetric and non-symmetric matrix projection cases in PART II of our paper.

3 The Hegedűs-Bodócs (HB) class of biorthogonalization algorithms (S6)

In this section we consider the subclass S6 of the ABS class. The HB biorthogonalization algorithms was first published in [14]. Recently a more detailed paper in this topic was published [13]. The main results of this section is Theorem 8.30 of [1] which proves how the HB algorithms constitute a part of the ABS class.

Theorem 4. *Consider the HB recursions with basis vectors* s_i, q_i *satisfying condition*

$s_i^T S_i A Q q_i \neq 0$

for all possible i , where

$$S_i^T = I - \sum_{j=1}^{i-1} \frac{Au_j v_j^T}{v_j^T Au_j}$$

and

$$Q_i = I - \sum_{j=1}^{i-1} \frac{u_j v_j^T A}{v_j^T Au_j}$$

where

$v_j = S_j s_j$

and

$u_j = Q_j q_j$

for $j = 1,...,i-1$. *Consider the following parameter choices in the scaled ABS class:* $H_1 = I$, v_i *and* z_i *given by*

$v_i = S_i^T s_i$

$z_i = Q_i q_i$

and w_i *a multiple of* z_i. *Then these parameter choices are well defined and moreover the following identity is true*

$p_i = Q_i q_i$.

Note that

$H_i^T z_i = z_i$.

therefore, based on the theoretical results the reason of the multiplication z_i by the projection matrix H_i^T is to have the possibility of the reprojections. As we show in our next paper the reprojection gives much better accuracy for the HB conjugate algorithms too.

It is important to note, that in this paper we suppose that the matrix A is positive definite symmetric matrix, consequently $p_i = Q_i q_i = S_i^T s_i$ that is the arbitrary vectors $v_i = p_i$ are defined as in the previous section. It means that Theorem 2 is valid for the Subclass S6 too.

Note also that the vectors z_i are still arbitrary.

In all algorithms listed below we also inserted the ABS versions to simplify the implementation. Many different versions of the HB algorithms follow from Theorem 8.30 of [1]. In the following definitions, for the sake of brevity, we leave out the index i wherever it is possible.

Algorithm p=H_ABS(v,u,Repr) (p is the conjugate direction)

$ABSv = v$

$ABSz = u$

$ABSw = ABSz$

$p = H^T * ABSz$

$s = HA^T p$

$if\ abs(s) < 3eps\ then$ % linear dependency

$\quad disp('the\ matrix\ A\ is\ singular')$

$\quad stop$

$endif$

$if \quad Repr == 1\ \ then$ %Reprojection is needed if Repr equals to one

$\quad p = H^T p$

end

$ptp = ABSv * Ap$

$pp = p/ptp$

$H = H - \frac{HA^T * ABSv * p^T}{p^T * A^T * ABSv}$.

Now we consider the following cases:

A) Hestenes–Stiefel algorithm in S6 (HBHSABS). The algorithm is defined by formulas (8.124) , (8.125), and the vectors s_i and q_i are defined by (8.135) and (8.136) in [1].

Algorithm P=H_HS_ABS(A,b,Repr,ReprHB,HB)

where A, b define the linear system, $Repr, ReprHB$ and HB are control parameters, see below.

Step 1 Initialize: Choose $S_1 = Q_1 = C_1 = K_1 = E$ where E is the n-dimensional unit matrix.

Let $\nu = \tau = 1$ be.

Compute

$r_1 = b - A * x$;

$s_1 = r_1$;

$q_1 = r_1$;

Step 2 (cycle for the dimension)

for i=1,...,n

$v_i = S_i^T s_i; \qquad u_i = Q_i q_i$

if $ReprHB == 1$ (Reprojection if $ReprHB$ equals to one)

$v_i = S_i^T v_i \qquad u_i = Q_i u_i$

endif

if $HB == 1$ (use the original version of the HS method in [13]

$P(:,i) = \frac{u_i}{norm(u,2)}$ (store the conjugate direction vector)

else

call **p=H_ABS(v,u,Repr)**

$P(:,i) = \frac{p_i}{norm(p,2)}$ (store the conjugate direction vector)

endif.

Step 3 Compute S_{i+1}, and Q_{i+1} by

$S_{i+1} = S_i - \frac{Au_i * v_i^T}{v_i^T Au_i} \qquad Q_{i+1} = Q_i - \frac{u_i * v_i^T A}{v_i^T Au_i}$

Compute the next arbitrary s_{i+1} and q_{i+1} vectors

$s_{i+1} = s_i - \frac{\mu_i s_i^T Cs_i}{v_i^T Au_i} Au_i \qquad q_{i+1} = q_i - \frac{\tau_i q_i^T Kq_i}{v_i^T Au_i} A^T v_i$

endfor.

B) Version of the HS method (S6CioccoHSDM). The algorithm is defined by formulas (3.3), (3.4) and (2.15) of [13].

Algorithm P=H_HSDM_ABS(A,b,Repr,HB)

Step 1 Initialize: Choose the positive definite Hermitian matrices $C = K = E$ as preconditioners where E is the n-dimensional unit matrix. Let x be an arbitrary vector which is not the solution of the linear system of equations. As C and K are unit matrices they are omitted from the formulas below.

Compute

$r_1 = b - A * x;$

$v_1 = r_1$

$u_1 = r_1$

$q_1 = r_1;$

$x = x + \frac{v_1^T r_1}{v_1^T Au_1} u_1.$

Step 2

for i= 1 : n

if $HB == 1$ (use the original version of the HS method in [13]

$P(:,i) = \frac{u_i}{norm(u_i,2)}$ (store the conjugate direction vector)

else

call **p=H_ABS(v,u,Repr)**

$P(:,i) = \frac{p_i}{norm(p_i,2)}$ (store the conjugate direction vector)

endif

$r_{i+1} = r_i - \frac{r_{i1}^T * r_i}{v_i^T A u_i} A u_i$ $\qquad$ $q_{i+1} = q_i - \frac{q_i^T * q_i}{v_i^T A u_i} A^T v_i$

$v_{i+1} = r_{i+1} + \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i} v_i$ $\qquad$ $u_{i+1} = q_{i+1} + \frac{q_{i+1}^T q_{i+1}}{q_i^T * q_i} u_i$

$x = x + \frac{v_{i+1}^T * r_{i+1}}{v_{i+1}^T A u_{i+1}} u_{i+1}$

endfor.

The next algorithm is an alternative numerical formulation of the previous one that is of H_HSDM_ABS. It is defined by formulas (2.2), (3.1) and (3.2) of (S6CioccoHSDMM).

Algorithm P=H_HSDMM_ABS(A,b,ReprHB,Repr,HB)

Step 1 Initialize: Define $PL = E$ and $PR = E$ where E is the n-dimensional unit matrix. Let x be an arbitrary vector which is not a solution of the linear system of equations. Compute

$r = b - A * x$

$rABS = -r$

$q = r$;

Step 2 (cycle for the dimension)

for i = 1 : n

$r = PLr$ $\qquad$ $q = PR^T q$

$v = PL^T r$ $\qquad$ $u = PRq$

if $ReprHB == 1$

$v = PL^T v$ $\qquad$ $u = PRu$

end

if $HB == 1$ (use the original version of the HS method in [13]

$P(:,i) = \frac{u_i}{norm(u_i,2)}$ (store the conjugate direction vector)

else

call **p=H_ABS(v,u,Repr)**

$P(:,i) = \frac{p_i}{norm(p_i,2)}$ (store the conjugate direction vector)

endif.

Step 3 update the matrices

$PL = PL - \frac{Auv^T}{v^T Au}$ $PR = PR - \frac{uv^T A}{v^T Au}$

end.

Note that the difference between the two algorithms from above is the reorthogonalization possibility in the second one. We shall have better accuracy in the solution with this reorthogonalization.

C) Lánczos type recursion in HB (S6CioccoLancz). The algorithm is defined by formulas (8.124) , (8.125), and the vectors s_i and q_i are defined by (8.139) and (8.140) in [1]. It is enough to define the basis vectors.

Algorithm H_Lánczos_ABS(A,b,Repr,HB)

Step 1 Initialize: Choose $S_1 = Q_1 = C_1 = K_1 = E$ where E is the n-dimensional unit matrix. As C_1 and K_1 are unit matrices they are omitted from the formulas below.

Let $\nu = \tau = 1$ be. Similarly we omit *nu* and τ from the formulas.

Compute

$r_1 = b - A * x$;

$s_1 = r_1$;

$q_1 = r_1$.

Step 2 (cycle for the dimension)

for i=1,...,n

$v_i = S_i^T s_i$; $u_i = Q_i q_i$

if $ReprHB == 1$ (reprojection if $ReprHB$ equals to one)

$v_i = S_i^T v_i$ $u_i = Q_i u_i$

endif

if $HB == 1$ (use the original version of the HS method in [13]

$P(:,i) = \frac{u_i}{norm(u,2)}$ (store the conjugate direction vector)

else

call **p=H_ABS(v,u,Repr,)**

$P(:,i) = \frac{p_i}{norm(p,2)}$ (store the conjugate direction vector)

endif.

Step 3 Compute S_{i+1}, and Q_{i+1} by

$S_{i+1} = S_i - \frac{Au_i * v_i^T}{v_i^T Au_i}$ $Q_{i+1} = Q_i - \frac{u_i * v_i^T A}{v_i^T Au_i}$

s_{i+1}, and q_{i+1} by

$$s_{i+1} = s_i - \frac{s_i^T q_i}{v_i^T A u_i} A^T v_i \qquad q_{i+1} = q_i - \frac{q_i^T q_i}{v_i^T A u_i} A u_i$$

endfor.

D) Method (S6Ciocco HSRM) defined by formulas (3.8), (3.9), (3.10) and (5.1) of [13]

Algorithm H_HSRM_ABS(A,b,Repr,HB)

Step 1 Initialize: Choose $PR = E, PQ = E$ where E is the n-dimensional unit matrix.

$v_1 = b - A * x$

$C = v_1^T v_1 E$

$K = v_1^T v_1 E.$

Step 2 (cycle for the dimension)

for k= 1 : n

if $k == 1$

$v_k = b - A * x$ $\qquad rr_k = v_k$

$u_k = v_k$

if $HB == 1$

$P(:,k) = u_i / norm(u_i, 2)$

endif

else

$u_k = PQ u_k$

if $HB == 1$

$P(:,k) = u_i / norm(u_i, 2)$

endif

endif.

Step 3

$x = x + \frac{v_k^T v_k}{v_k^T A u_k} u_k$

$\lambda_i = v_k^T * v_k$ $\qquad \varphi_i = \lambda_i$

$q_k^T = \frac{v_k^T APQ}{\varphi_i}$ $\qquad PQ = PQ - \frac{K * q_k q_k^T}{q_k^T * K * q_k}$

if $HB == 0$

call **p=H_ABS(v,u,Repr,)**

$P(:,k) = \frac{p_i}{norm(p,2)}$ (store the conjugate direction vector)

endif

endfor.

E) Method (S6Ciocco LDM) defined by (3.32), (3.33) and (5.1) of [13]

Algorithm H_LDM_ABS(A,b,Repr,HB)

Step 1

for k = 1 : n

if $k == 1$

$r = b - A * x$ $q = r$

$v = q$ $u = r$

$au = A * u$ $av = A^T * v$

$alp = v\prime * au$ $bet = q\prime * r; sig = bet/alp$

$x = x + sig * u$

if $HB == 1$

$P(:,k) = u/norm(u,2)$

else

call **p=H_ABS(v,u,Repr,)**

$P(:,k) = \frac{p_k}{norm(p,2)}$ (store the conjugate direction vector)

end

else.

Step 2 Preparation for the next iteration

$r = r - sig * au$ $q = q - sig * av$

$bn = q\prime * r$ $rat = bn/bet$

$v = q + rat * v$ $u = r + rat * u$

$au = A * u$ $av = A^T * v$

$alp = v\prime * au$ $bet = bn; sig = bet/alp$

$x = x + sig * u$

if $HB == 1$

$P(:,k) = u/norm(u,2)$

else

call **p=H_ABS(v,u,Repr,)**

$P(:,k) = \frac{p_k}{norm(p,2)}$ (store the conjugate direction vector)

endif

endfor.

F) Finally algorithm (S6HS) is defined by $ABSv_i = ABSu_i = ABSz_i = u_i$ where $ABSv_i$,$ABSu_i$ and $ABSz_i$ are the ABS class free parameter vectors.

Remark. In subclass S7 if we choose $v_i = Ar_i$ then the residual vectors r_i are A conjugate. The reprojection of the projection vectors does not give direct effect of the residual vectors. Therefore, we think that the accuracy of the solution would not grow very much. We present the test results of Subclass S6 and S7 in the third part of our paper.

4 Original algorithms

We implemented the original Hestenes–Stiefel and Lanczos methods as well.

1) Hestenes–Stiefel method (HSCGMoriginal). See in [15] or page 125 of [1].

Algorithm HS

Step 1 Initialize. Choose x_1. Compute $r_1 = Ax_1 - b$. Stop if $r_1 = 0$, otherwise set $p_1 = r_1$ and $i = 1$.

Step 2. Update x_i by

$$x_{i+1} = x_i - \frac{p_i^T r_i}{p_i^T A p_i}.$$

Step 3. Compute the residual r_{i+1}. Stop if $r_{i+1} = 0$.

Step 4 Compute the search vector p_{i+1} by

$$p_{i+1} = r_{i+1} - \frac{p_i^T A r_{i+1}}{p_i^T A p_i} p_i.$$

Step 5 Increment the index i by one and go to Step 2.

2) Lánczos method (Lanczosoriginal). See [18], [19] or page 126 of [1].

Algorithm Lanczos

Step 1. Initialize. Choose x_1. Compute $r_1 = Ax_1 - b$. Stop if $r_1 = 0$, otherwise set $p_1 = r_1$, $p_0 = 0$ and $i = 1$.

Step 2. Update the estimate of the solution by

$$x_{i+1} = x_i - \frac{p_i^T r_i}{p_i^T A p_i}.$$

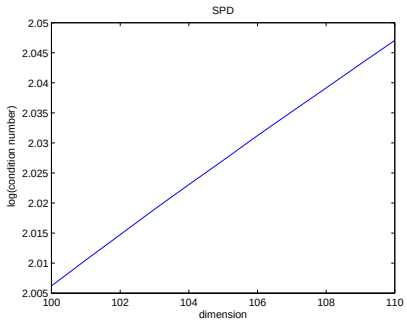
Step 3. Compute the residual r_{i+1}. Stop if $r_{i+1} = 0$. Step 4. Compute the search vector p_{i+1} by

$$p_{i+1} = Ap_i - \frac{p_i^T A^2 p_i}{p_i^T A p_i} p_i - \frac{p_{i-1}^T A p_i}{p_{i-1}^T A p_{i-1}} p_{i-1}$$

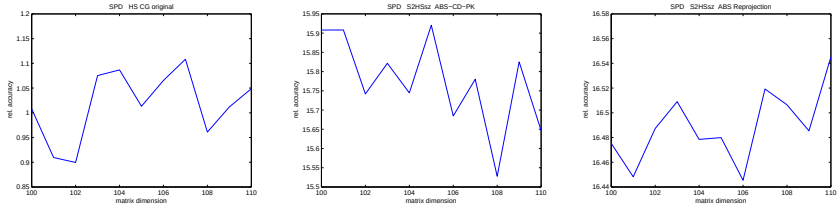
Step 5. Increment the index i by one and go to Step 2.

5 Preliminary test results

In this section we only show how the ABS_CD_PK algorithm works. We leave the intensive testing to the second and third part of the paper. To see the differences among the originals, the ABS_CD_PK conjugate directions method and the unconditional reprojection in the ABS methods we give an example. The algorithms were implemented in MATLAB version R2007b. The coefficient matrix is made by randomly generated Symmetric Positive Definite matrix (SPD) by the MATLAB rand function in [0,1]. Also the solutions of the constructed linear systems were generated randomly (by rand function) in [0,1]. The next figure shows the log(condition number) versus the considered dimension of the SPD problems



We chose the original Hestenes Stiefel, then the ABS_CD_PK algorithm with the Hestenes Stiefel method and the unconditional reprojection case in the ABS S2 subclass. The $\kappa = 100$ was chosen for the ABS_CD_PK algorithm which was suggested in [20]. The x axis shows the dimension while the y axis represents $y = -\log_{10}(yB)$ where $yB = \max abs(P^T AP - diag(P^T AP))/norm(A)$, where $norm(A)$ is the Frobenius norm of A.



where HS CG is the original Hestenes Stifel method see in [15] or page 125 of [20] for example. The name S2HSsz ($z_i = r_i$, $w_i = A^T p_i$) is the ABS symmetric version of it.

The norms of residuals in the solutions in case Hestenes Stiefel original are 3.826e-014 1.096e-013 6.628e-014 1.253e-013 6.889e-014 4.082e-014 7.418e-014 6.628e-

014 8.988e-014 5.64e-014 5.27e-014.

While in case S2HSsz ABS_CD_PK are 3.905e-013 4.353e-013 4.696e-013 4.187e-013 4.203e-013 4.264e-013 5.457e-013 4.942e-013 5.631e-013 6.169e-013 5.155e-013

The numbers of the reprojections are 31 35 44 38 38 41 41 35 49 38 44.

The numbers of linear dependency (ABS) are 10 17 19 22 10 16 10 8 17 14 13.

Finally in case of the unconditionally reprojecting method (ABS reprojection) the norms of residuals in the solutions are 4.092e-013 3.666e-013 5.04e-013 4.535e-013 4.49e-013 3.998e-013 5.749e-013 5.13e-013 4.951e-013 5.876e-013 5.498e-013

and the number of linear dependency (ABS) are 19 15 10 14 8 11 11 11 11 16 13.

The figures show the usefulness of the reprojection algorithm.

It can be seen that the S2HSsz ABS_CD_PK algorithm gives very good accurate results with much less number of computations than reprojections in every step.

Conclusion
In the paper we developed Parlett-Kahan's "twice is enough" algorithm for conjugate gradient case in the ABS class. The preliminary example shows the validity of our results. In the next two parts of the paper we intensively test many algorithms in the ABS class.

Acknowledgement
I would like to thank to my PhD student, Szabolcs Blága for his valuable help in writing this paper and to Csaba Hegedűs for the discussion on Subclass S6.

References

[1] Abaffy, J., and Spedicato, E., "*ABS Projection Algorithms: Mathematical Techniques for Linear and Nonlinear Equations"*, Ellis Horwood Limited, John Wiley and Sons, Chichester, England, (1989).

[2] Abaffy, J., Fodor, Sz., "*Reorthogonaliztaion methods in ABS classes*" under preparation

[3] Abdelmalek, N. N. "Round off error analysis for Gram-Schmidt method and solution of linear least squares problems" BIT 11 (1971) pp. 345-368

[4] Björck, A, "Solving linear least squares problems by Gram-Schmidt orthogonlaization" BIT 7 (1967) pp. 1-21

[5] Björck, A. and Paige, C. "Loss and recapture of orthogonality in the modified Gram-Schmidt algorithm", SIAM J. Matrix Anal. Appl. 13(1) (1992) pp. 176-190.

[6] Broyden, C.G. and Vespucci, M.T. "*Krylov Solvers for Linear Algebraic Systems*", Elsevier, (2004) ISBN 0-444-51474-0

[7] Galántai, A. and Hegedüs, C. J., "Jordan's principal angles in complex vector spaces", Numerical Linear Algebra with Applications 13 (2006) pp. 589-598 , http://dx.doi.org/10.1002/nla.491

[8] Giraud L., Langou J. and Rozloznik M. "On the round-off error analyisis of the Gram-Schmidt algrithm with reorthogonalization, CERFACS Technical Report No. TR/PA/02/33 (2002) pp. 1-11

[9] Giraud L., Langou J. and Rozloznik M. "The loss of orthogonality in the Gram-Schmidt orthogonalization process", Computers and Mathematics with Applications, Vol. 51 (2005) pp. 1069-1075,

[10] Golub, G. and van Loan, "*Matrix Computations*", 3rd ed. John Hopkins Univ. Press, Baltimore, MD (1966)

[11] Daniel, J.W, Gragg, W.B., Kaufman L. and Stewart G.W. "Reorthogonalization and Stable Algorithms for Updating the Gram-Scmidt QR Factorization", Mathematics of Computation Vol 30 No 136 (1976) pp. 772-795

[12] Dennis, J.E. JR and Turner, K. :"Generalized Conjugate Directions", Linear Algebra and its Application Vol 88/89 (1987) pp.187-209

[13] Hegedűs, C.J. : "Generation of Conjugate Directions for Arbitrary Matrices and Solution of Linear Systems", Proceedings of the NATO ASI Conference on Computer Algoirthms for Solving Linear Algebraic Systems, In Contributed papers of the NATO Advanced Study Institute Conference, Computer Algorithms for Solving Linear Algebraic Equations: The State of Art. (Sept. 9-22, 1990, Il Ciocco, Castelvecchio Pascoli, Tuscany, Italy.) (Eds. E. Spedicato and M. T. Vespucci), University of Bergamo, Bergamo, Italy, (1991) pp. 26-49.

[14] Hegedűs, C.J., and Bodócs, L." General Recursions for A-Conjugate Vector Pairs", Report No. 1982/56 Central Research Institute of Physiscs, Budapest (1982)

[15] Hestenes, M.R. and Stiefel, E.: "Methods of Conjugate Gradients for Solving Linear Systems" J. Res.Natlr. Bur. Stand. Vol 49 (1952) pp. 409-436

[16] Higham, N. J. *"Accuracy and Stabiility of Numerical Algorithms",* SIAM, Philadelphia, (1996)

[17] Hoffmann, W. "Iterative Algorithms for Gram_Schmidt orthogonalization" Computing Vol 41 (1989) pp. 335-348

[18] Lánczos, C. "An Iteration Method for the solution of the Eigenvalue Problem of Linear Differential and Integral Operators", J. Res.Natlr. Bur. Stand. Vol 45 (1950) pp. 255-282

[19] Lánczos, C. "Solution of Systems of Linear Equations by Minimized Iterations", J. Res.Natlr. Bur. Stand. Vol 49 (1952) pp. 33-53

[20] Parlett, B.N. "*The symmetric Eigenvalue Problem"*, Englewood Cliffs, N. J. Prentice-Hall (1980)

[21] Rice, J. R. "Experiments on Gram-Schmidt orthogonalization", Math. Comp. 20 (1966) pp. 325-328,

[22] Smoktunowicz, A. Barlow, J. L. and Langou, J. "A note on the error analysis of classical Gram-Schmidt", Numer. Math. 105/2, (2006) pp. 299-313,

[23] Wilkinson J. H. "*Rounding Errors in Algebraic Processes*", Prentice-Hall (1963)

[24] Wilkinson J. H. "*The Algebraiic Eigenvalue Problem",* Oxford University Press (1965)

[25] Zhang Liwei, Xia Zunquan and Feng Enmin, "*Introduction to ABS Methods in Optimization"*, Dalian University of Technology Press, (in chinese) (1998)

Runtime Translation of the Java Bytecode to OpenCL and GPU Execution of the Resulted Code

Razvan-Mihai Aciu, Horia Ciocarlie

Department of Computers, Faculty of Automation and Computers
Politehnica University Timisoara
Vasile Parvan Street, No. 2, 300223 Timisoara, Timis, Romania
razvan.aciu@cs.upt.ro, horia.ciocarlie@cs.upt.ro

Abstract: Modern GPUs provide considerable computation power, often in the range of teraflops. By using open standards such as OpenCL, which provide an abstraction layer over the GPUs physical characteristics, these can be employed to solve general computation tasks. Massively parallel algorithms used in domains such as simulation, graphics, artificial intelligence can greatly expand their application range. It is of great importance for an application to run parts of itself on GPUs and in this respect a number of issues such as OpenCL code generation, data serialization and synchronization between application and GPU must be observed. At the same time, the GPU limitations impose some limits on their applicability to general computation tasks, so an application must carefully assess what parts are suitable for this kind of execution. The computing resources must be abstracted and when possible these should be interchangeable without modifying the source code. We present a new algorithm and library which dynamically generates OpenCL code at runtime for parts of its application in order to run these parts on GPU. Our library automatically handles tasks such as data serialization and synchronization. The practical results are significant and we succeeded in obtaining important speedups using only a straightforward Java implementation of the test algorithm, without any platform specific constructs.

Keywords: Java; OpenCL; GPU; code generation

1 Introduction

For massively parallel algorithms, GPUs can offer an important speedup, often reducing their execution time by several times. In bioinformatics, using highly optimized libraries and GPU finely tuned algorithms, speedups of up to 1000x were reported [1]. Two top consumer GPUs in 2015 are AMD Radeon™ Fury X [2] with 4096 streaming cores, 4 GB memory, 8.6 TFLOPS FP32 and NVIDIA GeForce® GTX™ Titan X [3] with 3072 streaming processors, 12 GB memory and 7 TFLOPS FP32.

These GPUs are optimized especially by FP32 computing, so their FP64 performance is much lower (GeForce® GTX™ Titan X has 0.2 TFLOPS FP64), and Intel Xeon X7560 has 72.51 GFLOPS FP64 [4]. From these data it can be seen that the GPUs are valuable computing resources and their use would greatly enhance certain applications. Especially if FP32 precision is sufficient and if the algorithm is highly parallel, a single GPU can offer the performance of many desktop CPUs.

Taking into account the above considerations, it is understandable that many researchers try to develop new algorithms and frameworks capable of employing the GPU computational power. In this respect two main technologies are dominant: OpenCL and NVIDIA CUDA. We will concentrate on the OpenCL approaches, because it is a vendor neutral, open standard supported by all major vendors. Many of the aspects discussed also apply to CUDA because the physical structure of different GPUs has many common elements and both OpenCL and CUDA are abstraction layers over that physical structure.

Up to OpenCL 2.1 [5], which provides both a high level language (a subset of C++14) and a portable intermediate representation (SPIR-V), OpenCL programs were restricted to a subset of ISO C99. For maximal performance, the application interface (API) is also provided in C. The creation of an OpenCL application mainly involves the following tasks:

- creating the OpenCL code which will run on GPU,
- conversion of the application data in a format suitable for GPU execution,
- application-GPU communication and synchronization,
- data retrieval from GPU and its conversion into the application format.

If we consider a direct translation of the application code and its associated data structures to OpenCL, it becomes apparent that the above tasks are in fact standard procedures which can be automatically addressed by specialized tools or libraries. For example, the translation of the application code into OpenCL code is in fact a problem addressed in compiler theory by translators which outputs code in a high level language. The other tasks can also be automated. The resulted translation is similar with the original, in the same way a compiler creates a new representation of the original code. We do not address here the problem of creating GPU optimized code from a general algorithm, but some situations can still be optimized. A general optimization is the use of intrinsics, directly translated into native constructions for specific situations. We propose an algorithm and a library which automatically handles the above tasks, greatly simplifying the GPU/application interoperation. Our implementation employs the mentioned optimizations, translating when possible to native OpenCL operations.

The automatic translation from application to OpenCL has the advantage that it is easy to use and that it hides most of the GPU related aspects. The programmer does not need to create special data structures and he does not need explicit

serialization, deserialization and synchronization in order to communicate with the GPU. Optimized kernels or libraries can also be used if needed, or they can replace in time the automatic generated code.

2 Related Work in OpenCL Code Generation

In this section we discuss some existing libraries which automatically convert parts of their application code to OpenCL. There are several approaches in doing this and we highlight some of the benefits and drawbacks of each method. The first method is to use the application source code to generate OpenCL, by employing preprocessor commands. This approach is taken by libraries such as Bolt [6] for C++. Bolt offers macros such as BOLT_FUNCTOR which take a part of the source code and transform it to a string representation. These strings are later combined to form an OpenCL kernel. There is no further processing of the generated strings and they are used in the form they were captured, linked together with some glue code and ordered in a certain form. This method is directly applicable to languages which are supersets of the OpenCL itself, such as C/C++. There are notable advantages, one of which we should mention is the fact that the data structures have the same layout both in the host application and in the OpenCL code. The exact data layout can be enforced by using alignment specifiers. This simplifies considerably the interoperability between the host and the kernel. If the application data has a compact structure (without pointers to separately allocated structures), it can be sent in its native form to GPU so there is no serialization/deserialization overhead. At the same time, the code is the same in the application and in kernel, which simplifies debugging and CPU fallback computation when no GPU is available. This method has some weak points, among, which is the necessity for all the GPU involved code and its dependencies to adhere to the OpenCL subset. For simple kernels this is simple to accomplish, but if there are library dependencies, the entire libraries must be written in an OpenCL compatible C/C++ subset. In the source code, all the dependencies which are not included as separate files must be explicitly enclosed in specific macros. Another disadvantage is that the kernels cannot handle runtime provided code, such as plugins or short code snippets or formulas that need to be run on GPU.

A second method is for the programmer to explicitly build at runtime a representation for the needed computation and to generate the kernel from that representation. This intermediate representation can be a form of Abstract Syntax Tree (AST). By using well known algorithms, code can be generated from this AST. This approach is taken by libraries such as ArrayFire [7]. The method is suitable especially to combine library provided GPU accelerated functions into single kernels, eliminating host-GPU transfers between these functions, as in the case they were run separately. The AST leaves are the data involved in

computation, interfaced by proxy adapters. The AST nodes are functions and operators. Using operators overloading, in many cases the AST building for expressions can be syntactically abstracted as expressions written using the common infix operators, including their precedence rules and parenthesis. The method is very flexible and if the AST nodes allow (if there are nodes for loops, declarations, etc.), any construction can be generated. Runtime provided expressions or code snippets can be easily compiled by simply parsing them and constructing the associated AST. There are some drawbacks of this method such as the need to manually write the AST. For small kernels, composed from simple expression which only combine library functions, this is an easy task, but for large kernels that need declarations, decisions and loops, this task can be complex and tedious. Another disadvantage is that the AST code is completely distinct from the application code and this makes it difficult to debug and to interface with the application structures, even using predefined adapters.

Another method is to use the reflection capabilities of a programming language which enables the application to inspect its own code. This method is suitable for languages with strong reflection capabilities, such as Java and C#. Using this method, the application decompiles parts of its own code and translates them into OpenCL. The translation starts with the computation entry point (generally an interface with a method which defines the kernel body) and adds to the generated code all the required dependencies in a recursive manner. The associated data structures are also decompiled and translated. This approach is taken by libraries such as Aparapi [8]. Aparapi generates OpenCL code at runtime from an instance which implements the abstract class Kernel. The application overridden *run* method of this class is the OpenCL kernel entry point. Aparapi also serializes/deserializes the data needed for the kernel execution. This approach is very flexible and it succeeds in automating many GPU related tasks. At the same time, the GPU kernel is mirrored by the application code, which results in certain benefits such as ease of debugging (by debugging the application using standard Java tools) and the capacity to use CPU fallback computation when no GPU is available. Another benefit is the possibility of translating specific constructs into OpenCL optimized forms. Many of the Java Math class methods have corresponding OpenCL primitives and other commonly used primitives, such as the dot and cross products can be added. This method also allows plugins for GPU execution to be added at runtime. A drawback of these methods is the increased time needed to generate a kernel due to the overhead incurred by the code disassembly, but this can be alleviated by caching the generated kernel. In the particular case of Aparapi, for now it is designed as a thin layer over OpenCL, so it uses some specific functions such as *getGlobalId*. More important, it does not handle data structures which contain objects (the support is limited to single dimension arrays of primitive types), so complex data structures need to be handled manually by writing adapter code. Support for reference types is planned on certain architectures, such as the Heterogeneous System Architecture (HSA).

Another approach is the OpenJDK Sumatra [9] project with the primary goal to enable the Java applications to take advantage of the GPUs or other devices from the same class. The Sumatra project aims to convert specific constructs such as the Java 8 Stream API to an abstract representation (HSA Intermediate Language – HSAIL) from which it can be further converted to specific architectures, such as CPUs or GPUs. A difference from the above cases is that the Sumatra project tries to delegate the GPU interconnection tasks to the OpenJDK components (compiler, virtual machine), making the GPU a first class citizen of the Java supported architectures. For now the Sumatra project is in its early stages and it also depends on the success of emerging technologies such as HSA, but if it succeeds, it can be a significant achievement for the Java heterogeneous computing capabilities.

3 The Proposed Algorithm and Library

Our algorithm uses runtime reflection to access the application code, followed by disassembly, analysis and code generation steps in order to convert the relevant code to OpenCL. In the execution phase the required data is serialized automatically and transferred to GPU. The algorithm also handles the GPU synchronization and results retrieval, followed by a conversion into the application data structures. The main objective is to abstract as much as possible the GPU execution. Exactly the same code should run both on GPU and on CPU, in order to allow easy debugging and to provide CPU fallback execution when no GPU is available. Our library can generate OpenCL code for moderately complex situations such as data structures containing reference types (especially classes), exceptions handling and dynamic memory allocation. In this respect we depart essentially from libraries such as Aparapi, which are thin layers over OpenCL and in which many of the OpenCL requirements are explicitly exposed. Of course the abstraction layer can imply some performance loss. If more optimization is required, the programmer can translate some constructs into a more idiomatic code for GPU execution.

The entry point into the algorithm is a tasks scheduler which enqueues user tasks, executes them on GPU and retrieves the results. We extend [10] the standard Java thread pools with MapReduce [11] semantics and asynchronous, event driven receiving capabilities, similarly with the Node.js [12] non-blocking I/O. The library is designed in a way that allows distributed execution on many computing resources such as CPU, GPU or remote computers. When a result is received, it is passed to the *set* method of an object which implements the *Destination* interface. This method receives the resulted data and a task unique identifier such as an array index. A *Destination* can abstract over simple collections such as arrays and maps, or it can implement more complex processing, by immediately handling the data. In this way, if the specific application algorithm allows, the results can be directly used on arrival without storing them first. The scheduler starts running the

tasks as soon as they are added. In the end, after all the tasks are added, a synchronization method is called in order to ensure the end of all computations and the availability of the results.

A task is implemented as an object which implements the *Task* interface with *run* as its single method. A task encapsulates all its specific data. It is created on the application side and asynchronously enqueued for GPU execution with the scheduler *add* method. This method also associates the task with a unique identifier which will be used later when the result is processed. The task data is serialized and sent to GPU. After computation the result (returned by the *run* method) is received and converted to the application format, then sent to the scheduler associated destination. A generic use is given in Figure 1.

```
// Result is any user class; it encapsulates a task result
class ResultProcessor implements Destination<Integer, Result>{
        // set arguments: task unique identifier, received data
        @Override public void set(Integer id, Result ret){
                // process the received data
        }
}
class TaskProcessor implements Task<Integer, Result>{
        public TaskProcessor(/*input data*/){
                // task initialization on the host (application) side
        }
        @Override public Result run() throws Exception{
                // processes input data on GPU and returns it to application
        }
}
// ... entry point ...
// Scheduler instantiation
Scheduler <Integer, Result> scheduler=new Scheduler<>(new ResultProcessor());
// create all tasks and add them to scheduler
for(/*all data to be processed*/){
        TaskProcessor task=new TaskProcessor(/*specific initialization data*/);
        scheduler.add(id, task);   // asynchronous addition of the task to scheduler
}
scheduler.waitForAll();            // wait for all tasks to be completed
```

Figure 1

A generic use of the library

For simple cases such as the processing of Java standard collections, predefined destinations are provided so there is no need to implement the *Destination* interface. It can be seen that the processing resource is fully abstracted. Inside the *Task* implementation or on scheduler there are no OpenCL specific instructions,

but only general concepts are used, such as the task unique identifier. These allow a better representation of the application domain and a better abstraction of the computing resource. More specific instructions are given only if necessary, such as the case when multiple computing resources are present and a specific one needs to be selected.

When a new task is added to scheduler, first the scheduler checks if it already has the generated OpenCL code for that task. If the generated code exists, it is used else the task code is retrieved via reflection, disassembled and its corresponding OpenCL code is generated. The task instances are also serialized and in order to do this, their fields are investigated using reflection and the content is put in a memory buffer which will become the kernels global heap. Both code and data disassembly and serialization are recursive operations and the process ends when all the dependencies were processed. After the code is run on GPU, the heap is retrieved and its content is deserialized into the application data structures. This step updates in application the data modified by the kernel and it makes available the kernel allocated structures.

3.1 Data Serialization and Retrieval

The serialized data and the heap space for dynamic memory allocation are provided in a single buffer referred as the OpenCL global memory space or heap. All data references are converted in offsets in this heap. The kernel accesses data with a heap based indexing, similar with an array access. This method has a performance impact if the GPU native code does not support indexed access. In this case, a separate instruction must add the offset to the heap base pointer. Approaches such as the Sumatra project can use the shared virtual memory (SVM) feature from OpenCL 2.0, because they are tied to a specific Java Virtual Machine (JVM) implementation (OpenJDK), but in the general case JVM does not enforce a memory layout for many of its data structures, so we cannot directly use SVM.

The serialization algorithm uses the JVM primitive values as is. The only mention here is that the host and the GPU endianness must be observed. The reference values are split into two cases: class instances and arrays. In both cases the first member is a unique identifier for that class or array type. For class instances all the primitive values and references are added in the order given by reflection. Each class is implemented as an OpenCL structure and all the class instances are created and accessed through this structure. For arrays the array length follows and then the elements. The arrays for primitives are separate types and the arrays for references are implemented as a single type of array of *Object*, using type erasure. In order to ensure that the offsets of the instance members or array elements are known when the instance itself is serialized, these are serialized first, using a recursive process. The static members are implemented as a part of the global context and they also are stored on heap.

3.2 Code Generation

JVM has some features which are not available in OpenCL, such as dynamic allocation, exceptions throwing/catching, virtual method calls and recursion. These features need to be implemented as a library on top of the OpenCL primitives. Other features such as calling host functions for now are unavailable in OpenCL, so there is no method in doing this other than stopping the kernel, making that call from inside the application and restarting the kernel. This forbids the use of I/O functions such as the file or network API.

In the first translation step the JVM bytecode of each method is simulated linearly (without looping or branching) using a symbolic stack. A cell in this stack is an Abstract Syntax Tree (AST) node and each bytecode which influences the stack will combine these nodes. For example a numeric constant push bytecode creates an AST leaf for a constant value and an addition bytecode creates a new addition AST node from the top two AST nodes on stack. Other nodes are created or combined by instructions which do not operate on stack for example, the *goto* opcode. After this step we have a full AST for each method. We traverse this AST in order to generate the OpenCL code.

Memory allocation is a complex topic and it is particularly important for languages with automatic memory allocation, especially if there are no value semantics for classes, so all class instances need to be dynamically allocated (if the allocation is not optimized by the compiler). On massively threaded environments such as the ones provided by GPUs, where several thousands of tasks can run concurrently, special designed memory allocators are strongly required, or else they will become a performance bottleneck [13]. In our implementation we used a lightweight memory allocator, similar with [14], which is capable for now only to allocate data. This approach ensures a high allocation throughput (only one atomic operation is used to serialize the heap access) and no memory overhead for the allocated blocks, but it does not reclaim the unused memory, so the programmer must be careful with the number of allocations. For Java applications with intensive memory allocation this can be a problem and probably our allocator must be extended with full garbage collecting capabilities. On the programmer side, taking into account that the cases when GPU execution is required are especially the cases when a high performance is needed, simple methods can be used to minimize memory allocations and to optimize the application both for GPU and for CPU by reducing the pressure on the memory allocator. Our own interface library for accessing native OpenCL functionality reduces memory allocations by reusing the existent objects. In order to accomplish this, we devised the object based operations, such as the vectors addition to operate on the first argument instead of creating a new result vector. Because in OpenCL there are no global variables, each function needs to receive the global state (in our case the heap base address) as a supplementary argument. This incurs no overhead on the function call because the OpenCL compiler is required to inline all the functions so no code is generated for parameters passing, but only for their actual use.

Even in version 2.1 (which is based on a subset of C++), OpenCL does not provide exceptions handling. Even if a programmer does not throw exceptions in the code designed for GPU execution, these exceptions may appear from the memory allocator if it runs out of memory. Because there is no support for facilities such as stack unwinding, we implemented the exceptions as the functions return values. In this respect every function returns an integer value, which is an offset on heap. If that value is 0 (corresponding to a Java null pointer), no exception was generated. If an exception is generated, the exception object is allocated on heap and its index is returned. Every function call is guarded for non-zero return values and if this case happens, the enclosing function immediately exits, propagating further the received exception. If the exception is not handled, the kernel will be terminated and the exception object will be passed to application as a result for that particular task. An *OutOfMemoryError* object is preallocated for out of memory cases. Because the functions returns are used for exception propagation, if the function needs to return a value, it is returned through a supplementary reference parameter added to that function. This parameter holds the address of the variable which will receive the returned value. Using this parameter, the *return* instruction stores its expression into the provided address. We analyzed the GPU native code resulted from the OpenCL code in order to evaluate the performance impact of this design decision. Because the OpenCL compiler is required to inline all its functions, we saw that the pointer indirection used for storing the return value was simply optimized away and a direct store was used instead, so there is no performance loss using this model. In the same way, the guard for non-zero returned values and early enclosing function exit were propagated to the origin point (in our case the memory allocator) and an immediate kernel exit is performed if the subsequent code does not contain *try…catch* clauses, so this checking was also optimized away.

Some core Java classes such as *Object*, *Integer* and *Math* are treated as intrinsics. This allows a better code generation, optimized to use the OpenCL predefined functions. Many *Math* functions are already defined in OpenCL, so these can be translated directly into native functions. Other OpenCL available functions, such as the dot and cross products are provided in an auxiliary math library, which is also treated as an intrinsic. If the code is not executed on GPU, this library automatically uses a regular Java implementation for its functions. As an example of code generation, in Figure 2 we present the Java method we implemented to compute a Mandelbrot point and in Figure 3 (augmented with some comments and formatted to reduce the number of lines) we present its OpenCL generated code.

In order to be able to generate code for overloaded functions, we devised a new name mangling system because the JVM system uses characters which do not translate in valid C identifiers. We needed also to differentiate between functions with the same name from different classes, so in our system we append to a function name both its context (package and class name) and its signature.

```
int mandelbrot(float x,float y){
        final int ITERS=256;
        float xi=0,yi=0;
        float xtemp;
        int iteration;
        x=translate(x, 1.2501276f, 3.030971f, 0.31972017f, 0.34425741f);
        y=translate(y, -2.9956186f, 1.8466532f, 0.03119091f, 0.0572281593f);
        for(iteration=0;xi*xi+yi*yi<2*2 && iteration<ITERS;iteration++){
                xtemp=xi*xi-yi*yi+x;
                yi=2*xi*yi+y;
                xi=xtemp;
                }
        return iteration;
}
```

Figure 2
The original Java version of the Mandelbrot method

OpenCL does not have pointers to functions or other indirect calls. The virtual functions which regularly are implemented using virtual tables with pointers to the function implementation need to be implemented in other way. In our implementation the unique class identifier, which is the first member of any class structure, can be used for this purpose. This identifier is not an index in heap, but it is an index into a host maintained vector with the structures of the generated classes, so for a specific kernel it maintains its value regardless the actual data serialized in heap. Using this id the virtual functions can be implemented by testing the current object id in a *switch* instruction with dispatches for all possible situations (for all classes from a specific hierarchy which provide a specific implementation of that function). In the same way can be implemented dynamic dispatch for all the implemented interfaces. For now our generator does not fully implement dynamic dispatch (virtual methods), so we can use only the Java final classes, on which the specific called method can be inferred by the compiler.

Because the GPUs do not provide an execution stack, recursive functions are not implicitly supported [15]. The programmer must provide an iterative version of the algorithm or the recursion must be implemented using an explicit stack. In this version our library does not support code generation for recursive functions. We consider this a limitation which can be solved and we are trying to address it in the next version.

```
// idxtype is the integer type needed to access the heap vector (unsigned int)
// _g is a pointer to heap
// _l is a pointer to a location where the function return value will be stored
// return: 0 if no exceptions occured, or a heap index for the exception object
idxtype    tests_D_T3Work_D_mandelbrot_LP_FF_RP_I
                                        (_G _g, idxtype _this, float x, float y, int *_1){
int        ITERS, iteration, _TV3;              // _TV* are temporary variables
float      yi, xi, xtemp, _TV0, _TV1, _TV2;
idxtype _0;                  // used to test if exceptions occured and propagate them
ITERS=256;
xi=0.0; yi=0.0;
// each function call is guarded against exceptions
if((_0=tests_D_T3Work_D_translate_LP_FFFFF_RP_F(_g, _this, x,
          1.2501276,3.030971, 0.31972017,0.34425741, &_TV0))!=0)return _0;
x=_TV0;
if((_0=tests_D_T3Work_D_translate_LP_FFFFF_RP_F(_g, _this, y, -2.9956186,
          1.8466532, 0.03119091, 0.0572281593, &_TV0))!=0)return _0;
y=_TV0;
iteration=0;
goto _TMP172;
_TMP173:;
_TV0=xi*xi;_TV1=yi*yi;_TV2=_TV0-_TV1;_TV0=_TV2+x;xtemp=_TV0;
_TV0=2.0*xi;_TV1=_TV0*yi;_TV0=_TV1+y;yi=_TV0;
xi=xtemp;_TV3=iteration+1;iteration=_TV3;
_TMP172:;
_TV0=xi*xi;_TV1=yi*yi;_TV2=_TV0+_TV1;
// FCMPG is a macro which implement the JVM FCMPG opcode
_TV3=FCMPG((float)_TV2,(float)4.0);
_TV0=_TV3>=0;
if(_TV0)goto _TMP177;
_TV3=iteration<256;
if(_TV3)goto _TMP173;
_TMP177:;
*_1=iteration;               // setup the return value
return 0;                    // return without exceptions
}
```

Figure 3
The OpenCL generated code for the Mandelbrot method

4 Practical Results

In order to test our algorithm and library, we created a Java application which renders an image through ray casting. Primary rays are sent, without future reflections or refractions. The scene is composed of 1301 spheres. On each sphere the Mandelbrot fractal is applied as a procedural texture, so each pixel is directly computed when needed, without applying any precomputed bitmap. Besides the procedural texture, for each point is applied a simple illumination model which takes into account the angle of intersection between the ray and the sphere normal on the intersection point. Each horizontal line is considered a separate task (a work-item). The final result is shown in Figure 4. We used different Java features such as classes, members of reference types, static members. All calculations were done in FP32. Java Math library offers mostly FP64 operations, so we wrote a wrapper around these, to convert them to FP32. Because of this some operations such as *sin()* and *cos()* are executed as FP64 on CPU and FP32 on GPU. We allowed this because we assumed that in a real-world scenario when FP32 data is used, the programmer does not want to convert it to FP64 when using *sin()* and *cos()*, but if possible he will use FP32 operations. On the OpenCL side all these operations are translated into native functions. We did not use OpenCL or CPU specific features and the application ran without any modifications on CPU and on GPU. In this version our library uses for Java bytecode manipulation the ASM v5.0.3 library [16]. For standard OpenCL bindings we use JOCL v0.2.0-RC [17].

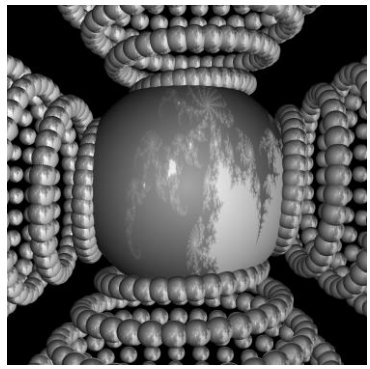


Figure 4
The result of the test program

We used the following test configuration:

- a computer with Intel® Core™ i5-3470 CPU at 3.20 GHz with 8 GB RAM, Windows 7 Home Premium SP1 on 64 bits and Java SE 8u45. This CPU has 4 cores.
- AMD Radeon™ R9 390X at 1060 MHz with 8 GB GDDR5 RAM. This GPU has 2816 streaming cores and 44 compute units.

We studied three main aspects: how the GPU execution compares with the CPU execution, how the GPU handles different workloads and how our library compares with the Aparapi library. For the comparison between GPU and CPU we used square images and we increased linearly the number of pixels in order to determine the best execution method for different sizes. Each test was run 5 times and the average value was taken. The GPU time included all the implied times: kernel generation and compilation, serialization/deserialization and the execution time. Data transfer between CPU and GPU are included in the execution time. This total time is needed only for the worst GPU case, when the computation is run only once. If the computation is run multiple times, the generated kernel can be cached and reused, so the generation and compilation time become insignificant. We measured these times across the test range and the results are given in Table 1.

Table 1

Setup times and data sizes for GPU

	10 KPixels image			43000 KPixels Image		
	Time (ms)	Heap data (KB)	Heap total (KB)	Time (ms)	Heap data (KB)	Heap total (KB)
Kernel generation	22	62.8	111	23.3	44442	45490
Kernel compilation	224			228.5		
Serialization	11.3			55.4		
Deserialization	0.65			100.5		

As expected, the kernel generation and compilation is invariant with the workload because the processed bytecode is the same. The serialization and deserialization times increase when more data need to be processed. For small workloads the result is given in Figure 5. Data was collected in increments of 10 KPixels (KP).

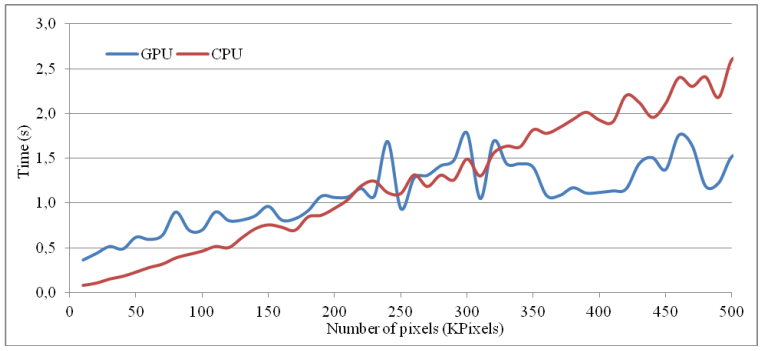


Figure 5

GPU vs. CPU execution for small workloads

It can be seen that the GPU execution starts to be faster at around 250 KP. If we subtract the time needed to generate and compile the kernel, this threshold lowers to around 100 KP. For this test application, the lower number of pixels is a GPU worst case because a square image of 100 KP has a height of around 316 pixels. It means that at maximum only 316 GPU streaming cores are working from the available 2816 streaming cores. Probably one of the few good things in this case is that a compute unit runs simultaneously fewer work-items, so the execution divergence is lower.

Next we compared the GPU vs. CPU execution across the entire possible range. On our system a single GPU execution is limited to about 30 seconds. After that, the operating system resets the graphic driver because it considers it unresponsive. On other systems this timeout can be lower, so care must be taken with long running kernels. The results are shown in Figure 6. Data was collected in increments of 1 MPixel.

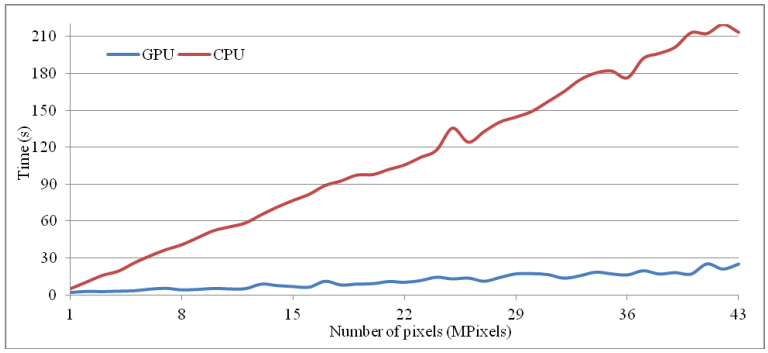


Figure 6
GPU vs. CPU execution across the entire test domain

It can be seen that when the quantity of work for a task (the image width) and also the number of tasks (the image height) increases, the GPU outruns the CPU with a linearly larger time. Because we increased the number of pixels linearly, the quantity of work also increases linearly. In this case, the difference in the number of the GPU streaming cores (2816) and the CPU cores (4) becomes apparent because in each case the increased workload is divided between all the available processing elements, so each CPU core will receive a greater amount of the increase. Even if both times increases are mostly linear, the CPU time increase is steeper than the GPU time. The maximum GPU speedup over CPU was 12.65x.

Another aspect we tested is how the GPU handles different types of workloads. For this test we kept the task size constant (the image width) and varied the number of tasks (the image height). We ran this test only on GPU and we measured the times for 3 different task sizes. The results are shown in Figure 7. Data was collected in increments of 64 tasks. We explain this graph by a combination of two factors. The general shape of each line is given by how the

workload fits in the GPU cache memory. It can be seen that for smaller workloads (width=1000), the increase is approximately linear over the entire domain. When we increase the workload, so more memory is needed, the number of cache misses increases and this increase becomes more apparent on the right side of the graph, where it strongly influences the time growth. The second factor which influences this graph is how the tasks are allocated for execution on GPU. If we add tasks so the total number is smaller than the number of the streaming cores, then the time increase for each added task is very small because all these tasks will be ran on the same batch. If the number of tasks is greater than the number of the streaming cores, we can have two extreme situations: all the tasks of a batch end approximately at the same time and in this case we have a local minimum, or when we increase the tasks by a small number the newly added tasks require a new batch only for them and in this case we have a local maximum. The situation is more complex due to the execution divergence and because not all the tasks require the same amount of work.

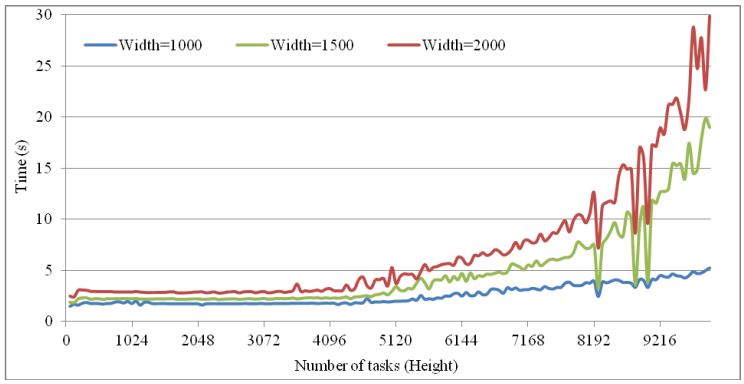


Figure 7
GPU execution for different number of tasks

A result of this analysis is that for long tasks it is better to reorganize them in such way that the amount of memory required by a task is as small as possible (or so they have common data). This optimizes the execution time by reducing the cache misses and also avoids the operating system timeout for GPU execution. The number of tasks sent for a single GPU execution can be set from the scheduler.

To compare our library with Aparapi, it was necessary to rewrite the test application into a representation suitable for Aparapi. We needed to replace some higher level data structures with lower level representations, such as:

- Aparapi does not support reference types (a part of unidimensional arrays of primitive types), so any classes used (such as Point, Line, …) were replaced with vectors of floats. For example, a Point was represented as a vector of 3 floats and a Line as a vector of 6 floats.

- Aparapi uses a single global space for all threads and the distinction between the threads data can be made using functions such as getGlobalId(). We needed to use combined vectors for all threads data and to access specific data based on the thread global id.
- Aparapi does not allow dynamic memory allocations so we needed to use global data structures to keep the returned functions values when these values are not primitive types. The thread global id was used to differentiate the threads data.

To illustrate some of these changes we show in Figure 8 how some parts of the test application are implemented using our library and in Figure 9 how these are implemented using Aparapi.

```
public class Point{
        public float x,y,z;
        public Point(){}
        public Point(float x,float y,float z){this.x=x;this.y=y;this.z=z;}
        public void set(float x,float y,float z){this.x=x;this.y=y;this.z=z;}
        public void set(Point p){x=p.x;y=p.y;z=p.z;}
        public float len(){return Math3D.length(x,y,z);}
        public void vectorFromPoints(Point origin,Point destination) {
                x=destination.x-origin.x;
                y=destination.y-origin.y;
                z=destination.z-origin.z;
        }
…
}
…
Point intersToRayOrigin=new Point();
Point centerToInters=new Point();
float     propagateRay(Line ray){…}
```

Figure 8
Part of the test application implemented with our library

Since the Aparapi version cannot use Java fundamental idioms such as classes, an algorithm written using this library needs additional proxy code to integrate with the rest of the application: translation from classes to vectors and back, merging of individual tasks data into single structures, etc.

In order to compare the execution of the Aparapi with our library, we varied both the number of pixels and the amount of work needed for each pixel. The latter measurement was needed in order to evaluate different workloads by keeping constant the amount of memory used and the number of threads. We recomputed each pixel a number of times (n), starting with n=1 (normal case). This process does not involve memory allocations.

```
final float pointLen(float []p){
        int offset = getGlobalId()*3;
        return length(p[offset],p[offset+1],p[offset+2]);
}
final void vectorFromPoints(float []dst,float []origin,float []destination){
        int offset = getGlobalId()*3;
        dst[offset]=destination[offset]-origin[offset];
        dst[offset+1]=destination[offset+1]-origin[offset+1];
        dst[offset+2]=destination[offset+2]-origin[offset+2];
}
…
final float []intersToRayOrigin;
final float []centerToInters;
final float        propagateRay(float []ray){…}
```

Figure 9
Part of the test application implemented for Aparapi

In Figure 10 we showed the results of the Aparapi executions and in Figure 11 the results from our library. For both figures data was collected in increments of 1 MPixel.

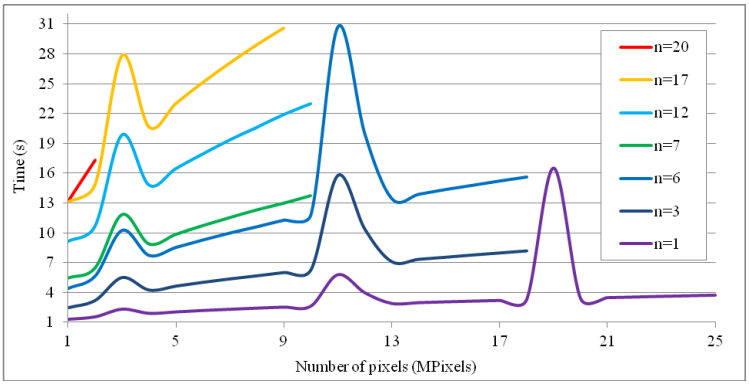


Figure 10
Aparapi execution for different number of tasks and recomputations

On the Aparapi version, the OS started to reset the graphic driver at around 25 MPixels. Our library was able to produce results up to 46 MPixels. Both libraries show approximately linear progressions, with some prominent peaks. Aparapi shows a better time and a smaller grow angle. Our library shows more irregularities from an ideal linear progression and, as discussed previously, we

consider this a combined effect of the data layout and the GPU cache. Aparapi has smaller irregularities because in its case the data are already vectorized and ordered by tasks, which improves the data locality. In both implementations the maximum running time, after which the OS started to reset the graphic driver was of about 31 seconds. When the peak would exceed 31 seconds, the application would crash.

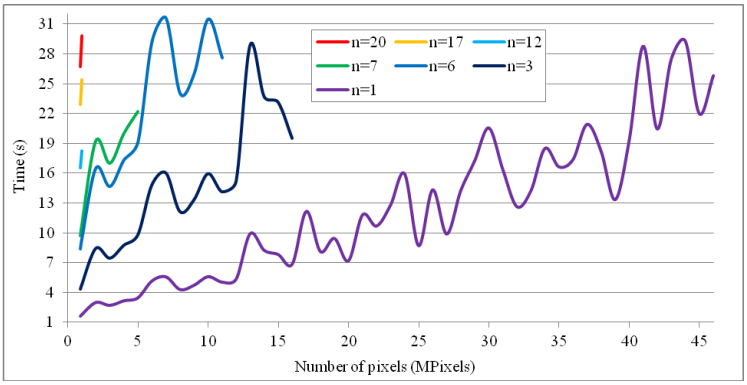


Figure 11
Our library execution for different number of tasks and recomputations

Conclusions

In this paper we proposed an algorithm and library which enable Java execution on GPUs. We used reflection in order to access the application code and OpenCL code generation to create the GPU kernels. This approach is also suitable for runtime provided code such as plugins. Our library automatically handles tasks such as data serialization/deserialization, GPU communication and synchronization. The data serialization system can process complex data structures, which enables the application to use for GPU execution classes, reference type fields, any type of arrays and static methods.

The library defines a Java compatibility layer over the OpenCL standard. This layer allows us to use exceptions handling and dynamic memory allocation. In the future we hope to extend it with virtual methods calls (dynamic dispatch) and recursive calls. Where possible, the OpenCL native functions are used instead of the Java standard libraries. We also provided an auxiliary library for the OpenCL primitives which do not have a correspondent into Java standard libraries. This library is portable, so it can be used both on GPU and on CPU.

Our algorithm uses a MapReduce model to manage the parallel tasks. The tasks return values are directly sent to a handler. In some cases this allows the results processing on arrival, without needing to store them. The tasks creation and management is abstracted over the computing resources, so the code can be executed both on GPU and on CPU without any modifications. This simplifies the

code maintenance, allows for an easy debugging of the code and the CPU can be used as a fallback resource when no suitable GPU is available.

We tested our library using a test application written in standard Java code, without any OpenCL specific constructs. For our test configuration we obtained significant speedups of up to 12.65x on GPU over the CPU execution. We consider the most important conclusion of this research the fact that parts of standard Java applications which use classes, dynamic memory allocation and exceptions handling (but for now without virtual calls and recursion) can be automatically translated to OpenCL and run on GPU and this can bring certain advantages. Our algorithm and library provides the capability to write Java code which can be easily integrated with complex data structures, code which does not need platform specific calls. This greatly simplifies the goal of running complex applications on different computing resources such as CPU or GPU. This is an important achievement over existing libraries such as Aparapi, which were designed as thin layers over OpenCL and requires the programmer to use specific OpenCL calls. In our tests Aparapi obtained a better time than our library, but it was capable of handling only a restricted domain of the test data and it required coding the application in a manner, which is not specific to Java (for example without classes) and which requires proxy code to translate the application data to a suitable representation and back. Our future research will concentrate on increasing the range of the applications which can run on GPU, develop better optimizations and obtain an increased reliability for GPU execution.

Acknowledgement

This work was partially supported by the strategic grant POSDRU/159/1.5/S/137070 (2014) of the Ministry of National Education, Romania, co-financed by the European Social Fund – Investing in People, within the Sectoral Operational Programme Human Resources Development 2007-2013.

References

[1] Lorenzo Dematté , Davide Prandi: GPU Computing for Systems Biology, Briefings in Bioinformatics, Vol. 11, No. 3, pp. 323-333, 2010

[2] AMD: A New Era in PC Gaming, E3, Los Angeles, California, U.S., 2015

[3] Jen-Hsun Huang: Opening Keynote, GPU Technology Conference, San Jose, California, U.S., 2015

[4] Paweł Gepner, David L. Fraser, Michał F. Kowalik, Kazimierz Waćkowski: Evaluating New Architectural Features of the Intel(R) Xeon(R) 7500 Processor for HPC Workloads, Computer Science, Vol 12, 2011

[5] Khronos Group: The Open Standard for Parallel Programming of Heterogeneous Systems, https://www.khronos.org/opencl/, download time: 14.07.2015

[6] AMD: Bolt, https://github.com/HSA-Libraries/Bolt, download time: 16.07.2015

[7] Kyle Spafford: ArrayFire: A Productive GPU Software Library for Defense and Intelligence Applications, GPU Technology Conference, San Jose, California, U.S., 2013

[8] AMD: Aparapi, https://github.com/aparapi/aparapi, download time: 16.07.2015

[9] Eric Caspole: OpenJDK Sumatra Project: Bringing the GPU to Java, AMD Developer Summit (APU13), 2013

[10] Razvan-Mihai Aciu, Horia Ciocarlie: Framework for the Distributed Computing of the Application Components, Proceedings of the Ninth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX, 2014, Brunów, Poland, Springer, ISBN: 978-3-319-07012-4

[11] Jeffrey Dean, Sanjay Ghemawat: MapReduce: Simplified Data Processing on Large Clusters, Commun. ACM 51, 2008

[12] Mike Cantelon, Marc Harter, T. J. Holowaychuk, Nathan Rajlich: Node.js in Action, Manning, 2014, ISBN 9781617290572

[13] Markus Steinberger, Michael Kenzel, Bernhard Kainz, Dieter Schmalstieg: ScatterAlloc: Massively Parallel Dynamic Memory Allocation for the GPU, Innovative Parallel Computing (InPar), San Jose, California, U.S., 2012

[14] Chuntao Hong, Dehao Chen, Wenguang Chen, Weimin Zheng, Haibo Lin: MapCG: Writing Parallel Program Portable between CPU and GPU, Proceedings of the 19th international conference on Parallel architectures and compilation techniques, PACT '10, Vienna, Austria, 2010

[15] Ke Yang, Bingsheng He, Qiong Luo, Pedro V. Sander, Jiaoying Shi: Stack-Based Parallel Recursion on Graphics Processors, ACM Sigplan Notices (Vol. 44, No. 4, pp. 299-300), 2009

[16] Eugene Kuleshov: Using the ASM Framework to Implement Common Java Bytecode Transformation Patterns, Sixth International Conference on Aspect-Oriented Software Development, Vancouver, British Columbia, Canada, 2007

[17] Java bindings for OpenCL, http://www.jocl.org/, download time: 19.07.2015

Models and Methods for Quality Management Based on Artificial Intelligence Applications

Nafissa Yussupova[1], George Kovács[2], Maxim Boyko[1], Diana Bogdanova[1]

[1] Faculty of Informatics and Robotics, Ufa State Aviation Technical University, 12 K. Marx str., 45000 Ufa, Russian Federation, e-mail: yussupova@ugatu.ac.ru, maxim.boyko@ugatu.ru, diana.bogdanova@ugatu.ru

[2] Computer and Automation Research Institute, Hungarian Academy of Sences, Kende u. 13-17, 1111 Budapest, Hungary, e-mail: kovacs.gyorgy@sztaki.mta.hu

Abstract: This paper proposes a conceptual approach to the research into customer satisfaction based on a detailed analysis of consumer reviews written in natural languages using Artificial Intelligence (AI) techniques such as Text Mining, Aspect Sentiment Analysis, Data Mining and Machine Learning. Special Internet resources for accumulating customer reviews, such as yelp.com, tripadviser.com and tophotels.ru, are used as data sources. To evaluate the efficacy of the proposed approach, we have carried out an experiment on qualitative and quantitative research of hotel client satisfaction. Even "Big data" applications were taken into account as a possibility to evaluate quality of services. The obtained results prove the effectiveness of the proposed approach to decision support in product quality management and argues for using it instead of classical methods of qualitative and quantitative research into customer satisfaction.

Keywords: quality management services; analysis of customer feedback; CPM; sentiment analysis

1 Introduction

Quality assurance is currently realized by means of a process approach based on the model of a quality management system [1]. It describes the interaction of the company and the customer during the process of product production and consumption. To correct the parameters of product quality in order to improve it for the customer, the model includes feedback. For companies, one aspect of feedback during the process of quality management is information about the level of customer satisfaction, expressed in the form of customer reviews of the product quality. That is why customer satisfaction is the key information in quality management that influences decision-making.

To collect data and to evaluate customer satisfaction, the International Quality Standard ISO 10004 recommends using the following methods: personal interviews, phone interviews, discussion groups, mail surveys (postal questionnaires), online research and survey (questionnaire survey) [2]. However, these methods of collecting and analyzing customer opinions show a number of significant drawbacks.

A general drawback of the recommended methods is the need for a large amount of manual work: preparing questions, creating a respondent database, mailing questionnaires and collecting results, conducting personal interviews, preparing a report based on the results. All this increases the research costs. Due to their discreteness these methods do not allow for the continuous monitoring of customer satisfaction. For this reason, the data analysis is limited to one time period and does not give an insight into the trends and dynamics of customer satisfaction. This also has a negative influence on the speed of managerial decision making, which depends on the arrival rate of up-to-date information about customer opinions.

Existing scales of customer satisfaction and their subjectivity perception raise additional questions. Values of customer satisfaction expressed in the form of abstract satisfaction indices make it difficult to understand, compare and interpret the results. Methods of analysis of data collected through the recommended ISO 10004 procedures permit only the detection of linear dependencies.

In this paper, to increase the effectiveness of product quality management, we suggest approaching the research of customer satisfaction through the use of AI technologies.

1.1 Should We Use Big Data for Analysis?

Recently several works deal with the problems and applications of “Big data”, as for example [3, 4 and 5]. Most often, the problem is associated with the necessity of processing of structured and unstructured data of large volumes. The term “Big data” appeared lately, in 2008 [6]. However, already in 2001 specialists in artificial intelligence faced the Big data problem, when a program to create "Intelligent Image" ended in failure. That time "Big data" was not identified as a separate problem, and has been regarded as a temporary difficulty. In general, the problem and the fact that the emergence of big data can be correlated, became clear only, when it was investigated from the approach of analytical data processing.

Big data is a series of approaches, tools and methods for handling structured and unstructured data of huge volume and significant diversity for results perceived by humans, effective in the conditions of continuous growth. The distribution of information across multiple nodes of computing networks in the pst 15 years gave the alternative to traditional control systems databases and decisions as a class Business Intelligence.

According to one of the approaches to this issue, the concept of "Big data" refers to the operations that can be performed only on a large scale.

Based on the realities of the world the use of "Big data" technology is becoming more and more important and commonly used. However, it is hard to define whether big data technology is really necessary and helpful for a given problem or not. The implementation of Big Data is guided by the concept of the four Vs: Volume, Variety, Velocity and Value [7]. To facilitate decision-making on implementing the technology "Big Data" in our case, we have estimated an “indicator of readiness” of transition to new technologies to work with a large volume of data - an indicator of readiness for Big Data is called Bigd. We applied the method, which is detailed in [8]. If the value of Bigd is more than 50%, the “t” technology of Big Data should be implemented. The Parameter «Volume» shows the size of the accumulated data, parameter «Velocity» is calculated from two values: the first describes the capture and processing of data in near-real-time (obtaining data by high-speed streaming); the second - is the rate of accumulation in the organization of data to be analyzed. If we estimate the increase of generated data to 60% or more per year, it results in unsolvable problems for companies with no appropriate IT infrastructure. The existing IT infrastructures would be soon exhousted and a necessary upgrade would cost much more than the benefits it would provide.

Parameter "Variety" is defined as follows: "the data is collected from one or more sources, and possibly in different formats." This parameter is determined by experts as the aggregate value of the quantitative sources. Parameter «Value» is determined by experts and is in the range from 0 to 1, and shows the value of the information from the source of data.

Our analysis showed that due to the features of our problems Big Data technologies cannot be effectively used to evaluate the quality of services. Therefore, to support decision making in the quality of services, management has proposed a method for evaluating the tone of reviews based on artificial intelligence technologies.

2 The Approach to Quality Management

Figure 1 represents the algorithm of the suggested approach to quality management based on research into customer satisfaction using AI applications. It consists of four main stages: 1. collection of reviews from Internet resources, data cleansing and loading data into the database. The second stage comprises the processing and analysis of the collected reviews. It includes marking reviews by their emotional response, i.e., sentiment (for example, negative and positive), identifying product aspects, and evaluating the sentiment of the comments on the

separate aspects. Following the stage of data processing utilizing visualization tools, quantitative research is carried out. A qualitative research of customer satisfaction is undertaken by means of building models based on decision trees, where the review's sentiment serves as a dependent variable, and sentiment comments on separate product aspects as independent variables. Managerial decision development and making is carried out on the basis of this research.

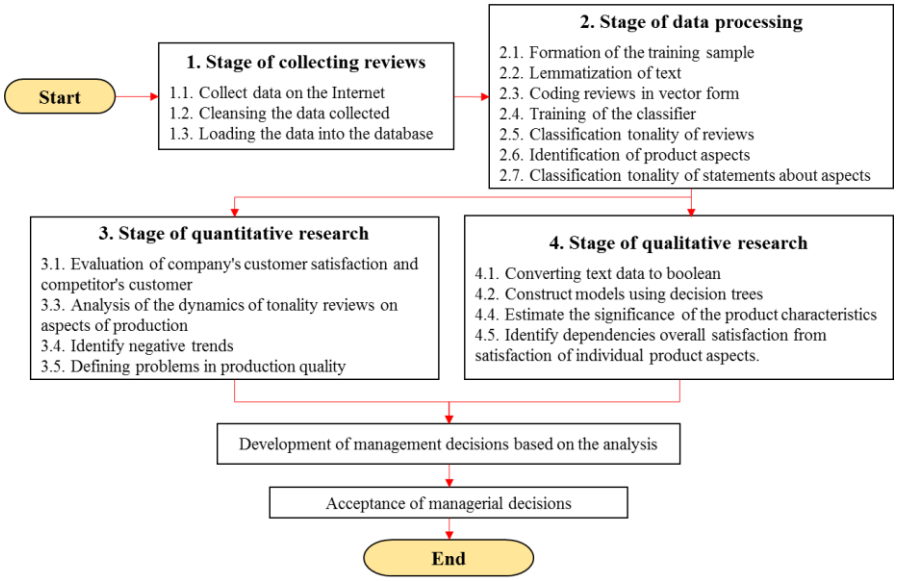


Figure 1
Quality management based on research of customer satisfaction using AI applications

3 Applied Artificial Intelligence Techniques

3.1 Data Collection

Nowadays there are a large number of Internet resources where users can leave their opinions about goods and services. The most popular examples are tophotels.ru (635,000 reviews), yelp.com (53 million reviews), tripadvisor.com (travels, 130 million reviews). Similar resources continue to gain popularity. Their advantage as a source of information for satisfaction evaluation lies in their purpose – the accumulation of customer reviews. As opposed to social network services, the web pages of review sources use XML that determines the structure typical for a review. Such a structure includes separate blocks with the name of a

product or company and a review, and other blocks with additional information. Therefore, all reviews are clearly identified in relation to the review object. It significantly simplifies the process of data collection and excludes the problem of key word ambiguity. One further advantage is that many of such resources monitor the reviews and check the objectivity of the authors.

There are two main types of collecting Internet data on customer reviews: 1) by using API (application programming interface) and 2) by web parsing. API is a set of ready-to-use tools – classes, procedures, functions – provided by the application (Internet resource) for use in an external software product. Unfortunately, only few resources that accumulate reviews have their own API. In this case, to collect reviews we can use the second method for data collection – web parsing. Web parsing is a process of automated analysis and content collection from xml-pages of any Internet resource using special programs or script.

3.2 Sentiment Analysis

After the data has been collected and cleaned, we can start their processing with the help of Text Mining tools. Sentiment Analysis is used to evaluate the author's product satisfaction. Sentiment stands for the emotional evaluation of an author's opinion in respect to the object that is referred to in the text.

We can distinguish three main approaches to Sentiment Analysis: 1) linguistic, 2) statistical, and 3) combined. The linguistic approach is based on using rules and sentiment vocabulary [9, 10 ,11]. It is quite time-consuming due to the need of compiling sentiment vocabularies, patterns and making rules for identifying sentiments. But the main drawback of the approach is the impossibility of obtaining a quantitative evaluation of the sentiment. The statistical approach is based on the methods of supervised and non-supervised machine learning. The combined approach refers to a combined use of the first two approaches.

The present work uses the methods of supervised machine learning: Bayesian classification and Support Vector Machines. Software implementation is simple, and does not require generating linguistic analyzers or sentiment vocabularies. Text sentiment evaluation can be expressed quantitatively. To apply these methods, a training sample was created. To describe an attribute space, vector representation of review texts was used with the help of the bag-of-words model. Bit vectors - presence or absence of the word in the review text, and frequency vectors – the number of times that a given word appears in the text of the review, served as attributes. Lemmatization, a procedure of reducing all the words of the review to their basic forms, was also used. More detailed information about machine learning methods used in this paper can be found in articles [15, 16].

3.3 Aspect Sentiment Analysis

Sentiment Analysis of reviews allows to evaluate general customer product or company satisfaction. However, it does not make clear what exactly the author of the review likes and what not. To answer this question, it is necessary to perform an Aspect Sentiment Analysis. An aspect means characteristics, attributes, qualities, properties that characterize the product, for example, a phone battery or delivery period, etc. However, one sentiment object can have a great number of aspects. Furthermore, aspects in the text can be expressed by synonyms (battery and accumulator). In such cases it is useful to combine aspects into aspect groups. An example of such aspect groups is represented in Figure 3.

An Aspect Sentiment Analysis of a review is a more difficult task and consists of two stages – identifying aspects and determining the sentiment of the comment on them. To complete the task of the Aspect Sentiment Analysis, a simple and effective algorithm has been developed:

First stage.

1. Extract all nouns S from the set of reviews D.

2. Count the frequency of words $\forall i = \overline{1,|S|} : f_i = N_i / N$ in the whole set of reviews D, where N is the number of appearances of all words, N_i the number of appearances of the i noun.

3. Count the difference $\forall i : \Delta_i = f_i - f_i^V$ between the counted frequencies f_i and *vocabulary* frequencies f_i^V.

4. Sort the set of nouns S in descending order Δ_i.

5. Divide the set of nouns *S* from $\Delta_i > 0$ into aspect groups.

Second stage.

1. Divide a set of reviews into sets of sentences.

2. Perform classification of sentiment for each sentence.

3. Check each sentence for the condition: if a sentence has a negative or positive sentiment and contains at least one noun from any aspect group, then the given sentence is labeled as an opinion about the given aspect.

A frequency vocabulary (based on the corpus) that helps to compare the obtained frequencies with word frequencies is used to identify aspects. The nouns with maximum frequency deviations are candidates for inclusion into aspect groups. Division of the noun set into aspect groups was carried out manually. We should note that if a sentence includes nouns from several aspect groups, then it will appear in each of them.

The results of Sentiment Analysis and Aspect Sentiment Analysis can be represented in the form of text variables $Obj = (\mathrm{Re}\, v_i, Sent_i, Neg_i^1, ..., Neg_i^j, Pos_i^1, ..., Pos_i^j)$, where Obj is a sentiment object or a product, $\mathrm{Re}\, v_i$ the text of the *i* review, $Date_i$ the date of *i* review publication, $Sent_i$ the sentiment of *i* review, Neg_i^j the negative sentences about the *j* aspect in the *i* review, Pos_i^j the positive sentences about the *j* aspect in the *i* review, *i* the review number, *j* the aspect group number.

3.4 Decision Trees

The following paragraph focuses on an algorithm of the processing of data obtained with help of Sentiment Analysis and Aspect Sentiment Analysis. The task of the developed algorithm is the mining of data that can be used for decision support in product quality management. To realize this algorithm, we use the intelligent data analysis tool, i.e. the decision tree since this tool can be easily understood and its results can be clearly interpreted; it also can explain situations by means of Boolean logic.

The algorithm of processing of data obtained by means of Sentiment Analysis consists of the following procedures:

1. Convert a set of text data $Obj = (\mathrm{Re}\, v_i, Sent_i, Neg_i^1, ..., Neg_i^j, Pos_i^1, ..., Pos_i^j)$ into a Boolean type by the following rules:

1.1. If $Sent_i = negative$, *then* $blSent_i = 1$ *else* $blSent_i = 0$.

1.2. If $Neg_i^j \neq null$, *then* $blNeg_i^j = 1$ *else* $blNeg_i^j = 0$.

1.3. If $Pos_i^j \neq null$, *then* $blPos_i^j = 1$ *else* $blPos_i^j = 0$.

2. Creating a decision tree where the variable $blSent_i$ is a dependent variable from $(Neg_i^1, ..., Neg_i^j, Pos_i^1, ..., Pos_i^j)$.

3. Estimation of the significance of aspect groups and interpretation of results.

The algorithm we have described allows us to understand which sentiment comments on product aspects influence the whole text sentiment or, in other words, what product aspects influence customer satisfaction and in what way. Our decision tree model allows us to consider the influence not only of separate sentiment comments on aspects but also of their mutual presence (or absence) in the text on customer satisfaction. The decision tree model also enables us to detect the most significant product aspects that are essential for the customer. The logical constructions (called rules) that we have obtained can be expressed both in the form of Boolean functions in a disjunctive normal form and in natural language.

A decision tree model can help to predict sentiment in dependence on various inputting aspect comments of different sentiments. In fact, it makes it possible to evaluate experimentally customer satisfaction in dependence on satisfaction with different product attributes. As the final result, prediction and analysis of the influence of different inputting variants on customer satisfaction allows us to distribute the company's budget effectively to maintain a high product quality.

The significance of aspects group shows how much the sentiment of a review depends on the sentiment of the aspect group. If the number of aspect groups is $g/2$, then the number of independent variables is g (positive and negative statements in each group of aspect). The formula for calculating the significance of m variable is:

$$Sign_m = \frac{\sum_{j=1}^{k_m}\left(E_{m,j} - \sum_{i=1}^{n_{m,j}} E_{m,j,i} \cdot \frac{N_{m,j,i}}{N_{m,j}}\right)}{\sum_{l=1}^{g}\sum_{j=1}^{k_l}\left(E_{l,j} - \sum_{i=1}^{n_{l,j}} E_{l,j,i} \cdot \frac{N_{l,j,i}}{N_{l,j}}\right)} \cdot 100\%, \quad (1)$$

where k_l is the number of nodes that were split by attribute l, $E_{l,j}$ is the entropy of the parent node, split by attribute l, $E_{l,j,i}$ is the subsite node for j, which was split by attribute l, $N_{l,j}$, $N_{l,j,i}$ are the number of examples in the corresponding nodes, $n_{l,j}$ is the number of child nodes for j parent node.

The score of customer satisfaction S with products is calculated by the formula:

$$S = \frac{N^{pos}}{N^{pos} + N^{neg}} \cdot 100\%, \quad (2)$$

where N^{pos} is the number of positive reviews, N^{neg} the number of negative reviews.

The score of customer satisfaction S_j with j aspect group of products is calculated by the formula:

$$S_j = \frac{n_j^{pos}}{n_j^{pos} + n_j^{neg}} \cdot 100\%, \quad (3)$$

where n_j^{pos} is the number of positive comments containing mention of the j aspect group, n_j^{neg} the number of negative comments containing mention of the j aspect group.

4 Experiments

Effectiveness evaluation of the developed approach was performed on the data obtained from 635,824 reviews of hotels and resorts in Russian. The reviews were collected from a popular Internet resource for the period of 2003-2013. The initial structure of the collected data consisted of the following fields: hotel name; country name; resort name; date of visit; opinion of the hotel; author evaluation of food; author evaluation of service; review number. The data was preliminarily processed and loaded into the database SQL Server 2012.

To classify segments, we used a binary scale (negative and positive) on the hypothesis that the absence of negative is positive. A training sample of positive and negative opinions was created using the collected data on the author's evaluation of accommodation, food and service. The Internet resource tophotels.ru uses a five-point grading scale. A review can have a maximum total of 15 points, a minimum of 3 points. The training sample included 15,790 negative reviews that had awarded 3 and 4 points, and 15,790 positive reviews that had awarded 15 points. We did not use author evaluation for further data processing. The marking of the remaining 604,244 reviews was carried out using a trained classifier.

For the purpose of effectively creating a sentiment classifier, we evaluated the accuracy of the classification of machine learning algorithms and some peculiarities of their structure (Table 1). The criterion Accuracy (ratio of the number of correctly classified examples to their total number) was used to assess classification accuracy. Accuracy evaluation was performed on two sets of data. The first set (Test No. 1) represented a training sample consisting of strong positive and strong negative opinions. It was tested by cross validation by dividing the data into 10 parts. The second set (Test No. 2) included reviews covering different points and was marked manually (497 positive and 126 negative reviews). It was used only for the accuracy control of the classifier that had been trained on the first data set.

To assess the influence of the negative particles "not" and "no", we used tagging; for example, the phrase "not good" was marked as "not_good", and was regarded by the classifier as one word. This technique allowed for the increase of sentiment classification accuracy.

Table 1
Comparison of methods for sentiment classification

Machine learning methods	**Vector**	**Test No. 1**	**Test No. 2**
SVM (linear kernel)	Frequency	94.2%	83.1%
SVM (linear kernel)	Binary	95.7%	84.1%
NB	Binary	96.1%	83.7%
NB	Frequency	97.6%	92.6%
NB (exceptional words)	Frequency	97.7%	92.7%

Bagging NB	Frequency	97.6%	92.8%
NB (tagging “not” and “no”)	Frequency	98.1%	93.6%

For the marking of reviews and the Sentiment Analysis, we created a classifier on the basis of the NB method, with frequency vectors as attribute space, and with the use of lemmatization and tagging of the negative particles “not” and ‘no”.

Using the algorithm we had developed we extracted from all reviews the key words that were divided into seven basic aspect groups (a part is represented in Figure 2): *beach/swimming pool, food, entertainment, place, room, service, transport.* The following step was extracting and marking sentences with words from aspect groups by sentiment. However, not all sentences with aspects have a clearly expressed sentiment; therefore, the sentences which do not show a clearly expressed sentiment were filtered out.

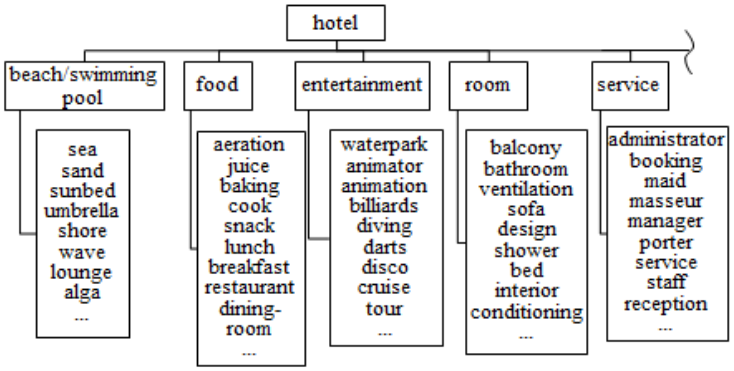


Figure 2

A part of the aspect groups of the object “hotel”

We will give an example of our qualitative and quantitative research for two 5 star hotels “A” (1,692 reviews) and “B” (1,300 reviews) located in the resort Sharm el-Sheikh (63,472 reviews) in Egypt. First, we will describe our quantitative research of consumer satisfaction dynamics, then we will compare this with the average satisfaction in the whole resort, detect negative trends in the different hotel aspects and identify problems in the quality of hotel services.

The dynamics of customer satisfaction is represented in Figure 3. Concerning Hotel “A”, there is a positive upward satisfaction trend beginning in 2009; it reaches the average resort level in 2013. Concerning Hotel “B”, in 2012 there was a sharp satisfaction decline and a similarly sharp increase in 2013. We can also notice this trend in a monthly schedule (Figure 4). Satisfaction decrease for Hotel “B” started in June 2012 and stopped in October 2012. Then, customer satisfaction with Hotel “B” grew to a level that was higher than the average resort level, being ahead of its competitor – Hotel “A”.

To find reasons for the Hotel "B" satisfaction decrease, we will examine the diagrams in Figure 5. We can see that in 2012, Hotel "B" on average was second to Hotel "A" in such aspects as "Room" (Δ12%), "Place" (Δ8%), "Service" (Δ5%), "Beach/swimming pool" (Δ3%) and "Entertainment" (Δ3%). Besides, in 2012, Hotel "B" had more registered cases of food poisoning as well as cases of theft in August 2012. We should also note that one of the reasons of client dissatisfaction with Hotel "B" as a place was the beginning of the renovation of the hotel building and the rooms. These measures, however, were rewarded in 2013, when customer satisfaction with Hotel "A" aspects equaled the average resort level.

In 2013, customer satisfaction with Hotel "B" exceeded the average level in all aspects (Figure 6). Customer satisfaction with Hotel "A" dropped to lower than average values in such aspects as "Service" (Δ3%), "Food" (Δ3%), "Beach/swimming pool" (Δ3%) and "Transport" (Δ4%). The manager of Hotel "A" could be advised to direct efforts to increase the quality of all aspects, but would this be the most effective solution? Which aspects are the most significant for the customer and should consequently be improved in the first place? Is it possible to offset the dissatisfaction with the service, for example, by healthier food or an animated evening performance and achieve client satisfaction? A qualitative research of the Sentiment Analysis results can give answers to these questions.

Figure 3
Dynamics of customer satisfaction by year

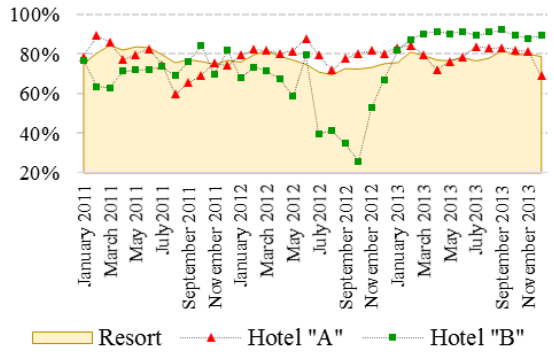


Figure 4
Dynamics of customer satisfaction by month

Decision trees were created using algorithm C4.5. The first step was creating a tree for the total sample of the reviews on the given resort to detect general principles. Extracted rules that have a reliability >80% are represented in Table 3. The second step is developing decision trees for the sample of Hotel "A" and Hotel "B" reviews to identify principles on the hotel level. Aspect significance is represented in Table 3.

Analyzing values of aspect significance (Table 6), we can say that the main factors of consumer dissatisfaction are a low service level, problems with food, and complaints about the hotel rooms. The most critical aspect for Hotel "B" is "Room". Without negative opinions on the aspect "Room", the reviews would be positive with a probability of 95.5% (Rule No. 10, Table 2). That is why the performed repair work facilitated a significant increase of consumer satisfaction. The most critical aspect for Hotel "A" is "Service", which corresponds with the findings for the resort as a whole.

The aspects which are significant both for the resort and for the two hotels and contributing to customer satisfaction are good food and amusing entertainment activities. The combination of these aspects can counterbalance negative emotions from the service or complaints concerning hotel rooms and leave the client with a favorable impression of the time spent in the hotel (Rules No. 5, 7, 11 Table 2). We should note that positive opinions about service, beach/swimming pool or place do not have a powerful influence on sentiment. That means the consumer priori focuses on a high-level of service, a well-kept place and the beach/swimming pool as a matter of course.

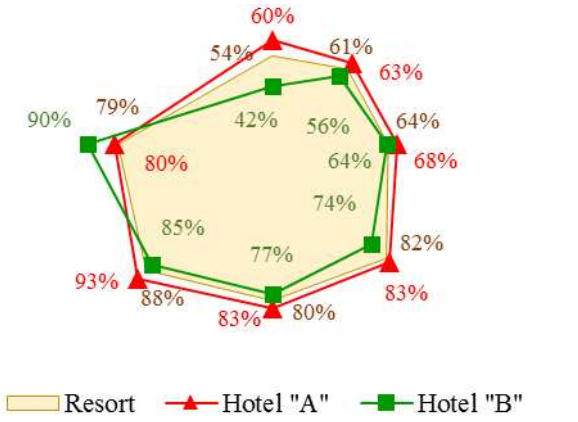


Figure 5
Comparison of customer satisfaction by aspects in 2012

The qualitative research we have undertaken enabled us to detect the main ways for Hotel "A" to increase customer satisfaction (Table 4). The problematic aspects identified in the course of our quantitative research correspond to the most significant aspects detected during the qualitative research stage. A search for

alternative aspects that can lead to customer satisfaction in the presence of negative opinions about the significant aspects “Service” and “Food” was carried out. To accomplish this, the rules (Table 3) containing negative sentiment in problem aspects, but which eventually lead to a positive review, were filtered out by the decision tree. The rules obtained and examples of appropriate managerial decisions are represented in Table 4.

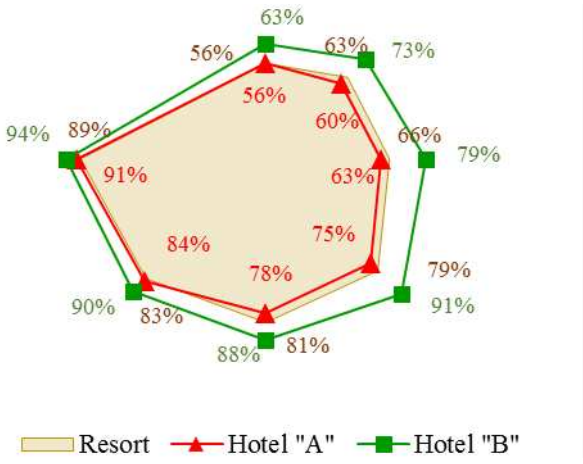


Figure 6
Comparison of customer satisfaction by aspects in 2013

Table 2
Significance of aspects

Aspect	Resort	Hotel “A”	Hotel “B”
Service$^-$	34.8%	60.2%	-
Food$^+$	30.3%	27.2%	30.3%
Food$^-$	16%	-	-
Entertainment$^+$	8.5%	12.7%	12.4%
Room$^-$	4%	-	57.3%
Beach$^+$	2.5%	-	-
Room$^+$	2.1%	-	-
Territory$^+$	1%	-	-
Service$^+$	0.7%	-	-
Beach$^-$	0.2%	-	-
Transport$^+$	-	-	-
Theft	-	-	-
Food poisoning	-	-	-
Entertainment$^-$	-	-	-
Territory$^-$	-	-	-
Transport$^-$	-	-	-

Table 3

Rules extracted by Decision Trees

№	Rules	Sentiment	Support	Reliability
Extracted rules on Resort reviews sample *				
1	$Food^{+} \cap \overline{Service^{-}} \cap \overline{Food^{-}}$	Positive	37-2%	97.4%
2	$Food^{+} \cap \overline{Service^{-}} \cap \overline{Food^{-}} \cap Beach^{+}$	Positive	11%	86.2%
3	$\overline{Food^{+}} \cap \overline{Service^{-}} \cap \overline{Service^{-}} \cap \overline{Room^{-}}$	Positive	10-6%	83.9%
4	$\overline{Food^{+}} \cap Service^{-} \cap \overline{Entertainment^{+}}$	Negative	6.9%	92.3%
5	$Food^{+} \cap Service^{-} \cap \overline{Food^{-}} \cap Entretainment^{+}$	Positive	5.8%	88.4%
6	$\overline{Service^{-}}$	Positive	62.9%	88.3%
7	$Food^{+} \cap Service^{-} \cap Entertainment^{+}$	Positive	20.5%	74.1%
8	$\overline{Food^{+}} \cap Service$	Negative	9.4%	86.2%
9	$Food^{+} \cap Service^{-} \cap \overline{Entertainment^{+}}$	Negative	7.2%	65.6%
* Accuracy of the created model by training sample 83.6% , by control sample 83.4%				
10	$\overline{Room^{-}}$	Positive	51.2%	95-5%
11	$Food^{+} \cap Room^{-} \cap Entertainment^{+}$	Positive	27.9%	81%
12	$\overline{Food^{+}} \cap Room^{-}$	Negative	11.1%	84%
13	$Food^{+} \cap Room^{-} \cap \overline{Entertainment^{+}}$	Negative	9.9%	55.8%

Table 4

Application of the results for the development of management solutions for Hotel "A"

Problematic Aspect Group	Rule level	Rules with result "Positive"	Support	Reliability	Examples of recommended managerial decisions
1. $Service^{-}$	Hotel	No.6: $\overline{Service^{-}}$	62.9%	88.3%	Educate and motivate service staff; check food service quality; organize entertainment activities.
		No.7: $Food^{+} \cap Service^{-} \cap Entertainment^{+}$	20.5%	74.1%	
	Resort	No.5: $Food^{+} \cap Service^{-} \cap \overline{Food^{-}} \cap Entertainment^{+}$	5.8%	88.4%	
2. $Food^{-}$	Hotel	-	-	-	Diversify menu; organize garbage collection on the beach.
	Resort	No.2: $Food^{+} \cap \overline{Service^{-}} \cap Food^{-} \cap Beach^{+}$	11%	86.2%	
3. $Beach^{-}$	Hotel	-	-	-	See above
	Resort	-	-	-	
4. $Transport^{-}$	Hotel	-	-	-	Not significant or outside of competence.
	Resort	-	-	-	

In order of preference, the manager of Hotel "A" should first of all make decisions on increasing the service quality, and then on increasing the quality of food and beach/swimming pool maintenance. Transport problems – concerning flights, early check-in, and baggage storage – are not significant and can be solved within the frames of service improvement. The process of service quality increase can take much time; that is why organizing entertainment and animated programs together with solving problems in connection with restaurant service and beach/swimming pool maintenance can serve as immediate measures to increase client satisfaction. Specification of managerial decisions can be performed on the basis of the information on existing problems contained in negative reviews. The extracted opinions on aspects can be used by hotel managers to improve specific service areas.

Conclusion

1) The suggested conception, based on the approach of text data processing and analysis that we have developed, allows us to undertake quantitative and qualitative research of customer satisfaction using computer-aided procedures and thus enabling the making of effective managerial decisions about product quality management. The present conception allows for the effective reduction of labor intensity of customer satisfaction research that makes it available for use by a wide range of companies.

2) The experiment performed has proved its effectiveness for solving real problems of quality management, a satisfactory accuracy of Text Mining algorithms, and consistency of the results obtained.

Future work in this research area can be devoted to the automatic annotation of text data, to the representation of the huge amount of text found in of the reviews in the form of a summary, and to extracting useful and unique information.

References

[1] ISO 9000:2008. The Quality Management System. Fundamentals and vocabulary

[2] ISO10004:2010. Quality Management. Customer satisfaction. Guidelines for monitoring and measuring

[3] Viktor Mayer-Schönberger, Kenneth Cukier Big Data: A Revolution That Will Transform How We Live, Work, and Think

[4] Chernyak L. Big Data - a New Theory and Practice // Open Systems. Database 2011, № 10, pp. 18-25

[5] Jacobs A. The Pathologies of Big Data // Communications of the ACM (2009) T. 52, № 8, pp. 36-44

[6] Lynch C. Big Data: How do Your Data Grow? // Nature (2008) T. 455, №. 7209, pp. 28-29

[7] Michael Minelli, Michele Chambers, Ambiga Dhiraj // Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses ISBN: 978-1-118-14760-3, p. 224

[8] Babaev E. O., Basha N. V., Tomsha P. P., St. Petersburg State University of Economics «Big Data» Defenition. Company Readyness Indicator to Implement New Techologies to Work with Big Data Ammount

[9] Yi J., Nasukawa T., Niblack W., Bunescu R. "Sentiment Analyzer: Extracting Sentiments about a Given Topic Using Natural Language Processing Techniques", In Proceedings of the 3rd IEEE international conference on data mining, ICDM 2003, pp. 427-434

[10] Pazelskaya A. G., Soloviev A. N. "Method of the Determination Emotions in the Lyrics in Russian", Computer Program Linguistics and Intellectual Technologies, Issue 10 (17), 201, pp. 510-522

[11] Ermakov A. E., Kiselev S. L. "Linguistic Model for the Computer Analysis of Key Media of Publications", Computational Linguistics and the Intellectual Technology: proceedings of the International Conference Dialog' 2005, 2005, pp. 172-177

[12] Pang B., Lee L. "Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval", Vol. 2, 2008, pp. 1-135

[13] Pang B., Lee L., "Thumbs up? Sentiment Classification using Machine Learning Techniques", Proceedings of the Conference on Empirical Methods in Natural. Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86

[14] Manning C., Raghavan P., Schuetze H. "An Introduction to Information Retrieval", Cambridge University Press. Cambridge, England, 2009, pp. 1-544

[15] Yussupova N., Bogdanova D., Boyko M. "Algorithms and Software for Sqentiment Analysis of Text Messages using Machine Learning", Vestnik USATU, T. 16-6(51), 2012, pp. 91-99

[16] Yussupova N., Bogdanova D., Boyko M. "Applying of Sentiment Analysis for Texts in Russian Based on Machine Learning Approach", IMMM 2012, Venice, Italy, 2012, pp. 8-14

Requirements Engineering in the Development Process of Web Systems: A Systematic Literature Review

José Alfonso Aguilar[1], Irene Garrigós[2], Jose-Norberto Mazón[2]

Universidad Autónoma de Sinaloa, Ave. Universidad y Leonismo Internacional S/N, 82140 Mazatlán, Sinaloa, México, ja.aguilar@uas.edu.mx

University of Alicante, Carretera San Vicente del Raspeig S/N, 03690 San Vicente del Raspeig, Alicante, Spain, {igarrigos, jnmazon}@dlsi.ua.es

Abstract: Requirements Engineering (RE) is the first phase in the software development process during which designers attempt to fully satisfy users' needs. Web Engineering (WE) methods should consider adapting RE to the Web's large and diverse user groups. The objective of this work is to classify the literature with regard to the RE applied in WE in order to obtain the current "state-of-the-art". The present work is based on the Systematic Literature Review (SLR) method proposed by Kitchenham; we have reviewed publications from ACM, IEEE, Science Direct, DBLP and World Wide Web. From a population of 3059 papers, we identified 14 primary studies, which provide information concerning RE when used in WE methods.

Keywords: Web Engineering; Systematic Literature Review; Requirements Engineering; Model-Driven Engineering

1 Introduction

A Web system (WS) is an application that is invoked with a Web browser over the Internet. This application has a set of special features, such as the inclusion of a multidisciplinary team for its development and its large and heterogeneous user community. Web systems are widely accessed by different types of users who have different needs, goals and preferences. These systems must additionally satisfy the needs of many types of stakeholders apart from the users themselves, e.g., the people who maintain the system, the organization that requests the system, or those who fund the system development budget. Special requirements must consequently be considered during the development of Web systems: (i) what information should be provided (content requirements), (ii) which scenarios should be defined in order to provide this information (navigational requirements), (iii) how the user or groups of users should be provided with this information and

(iv) how quality should be evaluated, which is in some cases encompassed in Non-Functional Requirements (NFRs) or named quality attributes, e.g., how to make Web content accessible to people with disabilities [1]. The development process therefore necessitates knowledge and expertise from many different disciplines and requires a team of diverse groups of people with a high degree of expertise in different areas [2], such as developers, designers and so on, thus making this process even more complex and difficult than normal software development. This has led to the appearance of WE methods such as OOH [3], WSDM [4], WebML [5] and UWE [6] which provide different mechanisms that can be used to consider the content, composition, and navigation features of Web systems, including the appropriate steps needed to consider requirements [7].

Bearing these considerations in mind, this paper presents a Systematic Literature Review (SLR) in order to analyze the current state-of-the-art with regard to Requirements Engineering (RE) in Web Engineering (WE), thus revealing the activities that are implemented, such as elicitation, analysis, specification, validation and management. An SLR is a means of identifying, evaluating and interpreting all the available research that is relevant to a particular research question, topic area or phenomenon of interest. It originated in the field of medical research and was successfully adapted to Software Engineering (SE) by Kitchenham [8].

In our previous work [10], we developed an initial SLR in which we began to highlight the importance of considering requirements during the development of Web systems. In the current SLR we have now improved many parts of the paper and have added a lot of new material. For example, one of the weak points of the previous SLR was the search strategy, and this has now been improved. More methods have consequently been added to the review (OOHDM [28] and HERA [38]). What is more, the study has been focused on analyzing the RE activities (elicitation, analysis, specification, validation and management) implemented in each of the WE methods investigated, how the requirements are dealt with as regards RE activities and how these methods have been released into the academic community, with special emphasis on their implementation in industrial projects and the tool support they offer for the entire development process and for RE. We have also analyzed the requirements terminology (terms used to name the special requirements for WS) adopted by each method in a more methodical and comprehensive manner based on the classification previously presented by Escalona and Koch [7].

The remainder of this paper is structured as follows: Section 2 presents those RE and WE concepts that are relevant to the context of this paper. The SLR is detailed in Section 3. The Research Questions are answered in Section 4, in which an analysis and discussion of this work and suggestions for future research are also presented. Finally, our conclusions are provided in Section 5.

2 Requirements and Web Engineering Concepts

Requirements Engineering (RE) is the process of discovering, analyzing, documenting and verifying the services that should be provided by software, along with its operational constraints [84]. Various approaches have been used to define RE activities, such as those proposed in [11, 12], and these activities widely differ from each other for several reasons, e.g., depending on the application domain, the people involved and the organization developing the requirements. However, there are a number of generic activities that commonly appear in all of them, such as elicitation, analysis, specification, validation and management. These are detailed as follows:

- Elicitation, whose goal is to discover what problems need to be solved [12], and to identify the stakeholders, and the objectives that a software system must attain. It is carried out through the application of various techniques [13, 14, 15], such as questionnaires, brainstorming, prototyping and modeling techniques, e.g., goal oriented based methods [16].
- Analysis, which includes the creation of conceptual models or prototypes with which to achieve the completeness of the requirements and deals with understanding an organization's structure, its business rules, goals and tasks, and the data that is needed. [84].
- Specification, which is an integral description of the behavior of the system to be developed. The most widely used techniques are templates, scenarios, use case modeling, and natural language [17].
- Validation. The aim of this phase is to establish whether the requirements elicited provide an accurate representation of the actual *stakeholder* requirements. Some of the techniques employed are reviews and traceability [18], [19].
- Management**,** which consists of recognizing changes through the use of continuous requirements elicitation, and includes techniques for configuration management and version control [20].

After this overview of RE concepts, it is worth noting that the development of WSs involves particular requirements that are different from traditional software requirements, as defined in the seminal work of Escalona and Koch [7], e.g., the authors put forward the argument that Functional Requirements (FRs) for WE are related to three main features of WSs: navigational structure, user interface and personalization capability. An overview of each kind of requirement is provided as follows: i) Content: This is the information that should be presented to users, e.g., in an online bookstore one example might be the information about a "book". ii) Service: The internal functionality with which users are provided. Following the online bookstore example, e.g.: "register a new client". iii) Navigational: The navigational paths the user can follow, e.g., user navigation from the "index page"

to different menu options such as “consult products by category”. iv) Layout: This defines the visual interface for users, such as “a color style”. v) Personalization: Personalization actions to be performed by the Web systems e.g. “show recommendations based on interest in previously acquired books” and vi) Non-Functional: These are related to quality criteria, e.g., “good browsing experience” and “improve efficiency”.

This classification of requirements is used throughout this SLR for the sake of understandability and completeness.

3 The Systematic Literature Review

The objective of this SLR is to summarize the information concerning how RE activities are applied in WE in order to detect avenues for future research.

3.1 Research Questions

According to [8], the question structure is divided into four aspects known as PICO (Population, Intervention, Comparison and Outcomes). The term Population refers to the people, projects and application types affected by the intervention. Intervention concerns the software technology, tool or procedure that generates the outcomes. Comparison refers to another type of intervention – if applicable – while Outcomes are the technological impact on relevant information terms for practical professionals. This PICO strategy has been used as the basis for our research and is use in this context is described as follows:

- Population: the population is composed of designers and developers who request a method in order to obtain more robust process support, and of researchers in the WE field who aim to develop new methods.
- Intervention: this review must search for indications that the development of WSs can be fully supported by a systematic process and a specialized tool.
- Comparison: not applicable.
- Outcomes: the objective is to demonstrate how a systematic process supports the development of WSs and whether or not the process is fully supported with regard to RE activities.

Our research questions (RQ), which are based on the aforementioned strategy, are:

RQ1. - Which of the existing methods for the development of Web systems are based on a systematic process? There are several WE methods for the development of WSs, but not all of them are necessarily based on a systematic method covering all the activities of the development process.

RQ2. -Which RE activities are supported by each method and the techniques used? The RE activities adapted to be applied in WE in each method are studied along with the techniques that they use, e.g., UML Use Cases.

RQ3. - Is there any common terminology with regard to RE that is applied by the existing methods? To detect the way in which the requirements are denominated by each WE method, e.g. "Functional Requirements" can be called "Service Requirements" and "NFRs" can be termed as "Softgoal's", and it is therefore necessary to establish a universal means of denominating Web requirements.

RQ4. – Which methods provide tool support for their development processes? To analyze those WE methods that provide tools covering the development process, including the RE activities.

3.2 Search Strategy

The search strategy should be systematic. According to [8, 9], it is necessary to use search engines by applying a combination of search terms (keywords) extracted from RQ's. Experts should then verify and review the search results. Once the steps to be followed in the search process have been defined, it is necessary to state the resources that are available to conduct the review of primary studies (individual studies contributing to an SLR). The research sources used are repositories with restricted access such as: ACM, IEEE, Science Direct, DBLP Computer Science Bibliography, World Wide Web: Google Scholar. In accordance with Brereton [22], these libraries were chosen because they are some of the most relevant sources in SE. Furthermore, Google Scholar was selected to complete the set of conferences and workshops searched and to seek grey literature in the field (white papers, PhD theses), and the results obtained were then compared with the works found using the search strings.

The structure of the research questions was used as a basis, to extract some keywords, which were then used to search for primary studies. We initially had the following keywords: *Web*, *engineering*, *requirements*, *development*, *method* and *tool*. However, in order to obtain more concrete and specific results in the field, we decided to link *Web* with the keywords *engineering* and *requirements*, *requirements* with the keyword *engineering*, and the keyword *Web* with the keywords *engineering* and *methods*. In this respect, the choice of concatenating "Web" with "engineering" was motivated by our goal, which was to retrieve papers specifically focused on RE in the Web domain. The search string "(Web OR WWW OR World-Wide Web OR Internet) AND engineering" was not therefore considered. The search string was used in all instances, even when examining papers from special issues on Web Engineering. Moreover, a list of synonyms was constructed for each of these keywords. Nevertheless, other words were also added in order to increase the size of potential relevant studies: *system*, *techniques*, *phase*, *and design*. These words were linked with the keywords *Web*, *requirements*, *methods*, *engineering* and *tool*. In order to avoid imprecise results,

we used the SLR proposed by Walia and Carver [23] and the review by Beecham [24] as a basis to create a specific type of string for each search engine (using the list of synonyms), thus making the search as accurate and comprehensive as possible. The search covered is the time period from 2009 to early 2015. This was initially 2014, but we then decided to extend the period to the first months of 2015 in order to obtain better results. Moreover, the corresponding authors of the main texts were e-mailed directly to clarify some particular issues regarding their Web methods, e.g. the authors of NDT and UWE. Finally, the references found in our primary studies and in renowned conferences such as the International Conference on Web Engineering (ICWE) and Web Information Systems and Technologies (WEBIST) were used to search for publications in order to ensure that no major works had been missed.

3.3 Inclusion and Exclusion Criteria

Essentially, only those publications from the RE literature regarding the development of WSs based on a method for specific use in the WE field were considered. Although our research questions are related only to WE methods, this SLR includes the primary studies related to the RE in the Web field and we therefore deemed that at least the part related to the use of one of the RE activities in WE must be present in each primary study, since we assumed that not all methods implement another RE phase. We chose the following inclusion criteria in order to select the relevant publications required to answer our research questions: i) Publication date between 01/01/2009 - 01/01/2015, ii) Requirements phase of WS development process, iii) Explicit mention of WE, iv) Relevance with regard to research questions and v) WE methods with tool support. The exclusion criteria were: i) Topics that do not match the RE activities implemented in Web methods and ii) Duplicated documents from the same study.

3.4 Study Quality Assessment

The place of publication and the diffusion of the methods were used as indicators when performing the quality assessment. The place of publication refers to the journals and conferences in which the primary studies were published (this applies to Google-Scholar which searches for a wider spectrum of papers such as white papers). The diffusion of the methods corresponds to the academic or industrial application of the method, including tool support and whether the tool is a prototype or an industrial-commercial tool. The first search, during which no exclusion criteria were applied, returned a total of 3059 documents of which 70 documents were duplicated. After applying the exclusion criteria (a further review round), 14 documents were eventually considered. It is important to mention that the activity during which publications were searched for was checked by two individual authors of this work in order to verify the quality of the place of publication. The quality assessment was then performed separately to verify the information extracted.

3.5 Data Extraction

The goal of this phase is to design data extraction forms with which to accurately record the information obtained from the primary studies. This form must be designed in such a way that all the information required can be collected in order to fully address the research questions. It was at this point of our SLR that the data extraction was performed. We used a form to store the information extracted from the search results, storing the publication title, the journal or conference/workshop in which the paper was published, the publication date, the main author, the RE techniques, the shortcomings with regard to RE and the tool support.

After quality assessment, the data synthesis was performed. This was done by collating and summarizing the results of the primary studies, which are: *Metamodeling the requirements of Web systems* [31]; *Model transformations from requirements to web system design* [32]; *Requirements engineering for Web Applications - a comparative study* [7]; *Introducing requirements traceability support in model-driven development of web applications* [33]; *The object-oriented hypermedia design model* [30]; *Integrating usability requirements that can be evaluated in design time into Model Driven Engineering of Web Information Systems* [34]; *From task-oriented to goal-oriented Web requirements analysis* [35]; *Transformation techniques in the Model-Driven Development Process of UWE* [36]; *NDT. A Model-Driven Approach for Web requirements* [27]; *A requirement Analysis Approach for Using i* in Web Engineering* [21]; *Web Modeling Language (WebML): a modeling language for designing Websites* [5]; *WSDM: a user centered design method for Web sites* [37]; *Hera: Development of semantic web information systems [38] and Extending a Conceptual Modeling Approach to Web Application Design* [39]. In this respect, it is important to mention that the synthesis of these primary studies was descriptive (non-quantitative) [8, 9] and was carried out by answering the RQ.

4 Data Analysis

This section presents and analyzes the results obtained after subjecting the primary studies to the extraction and data synthesis activities. The selected studies provided relevant evidence with which to satisfactorily answer the four RQs, as described below:

RQ1. - Which of the existing methods for the development of Web systems are based on a systematic process?

The methods extracted from the selected publications were OOWS [25], NDT [27], OOHDM [28, 29, 30], A-OOH [21, 40, 41, 42, 43], UWE [6, 32, 36, 56], WSDM [4, 37, 44], WebML [5, 45, 46, 47], and HERA [38, 48]. Since the A-OOH [54], OOWS [49, 50] and UWE [56] methods have a development process

that is based on Model-Driven Architecture (MDA), a three layer architecture, the process is considered from the first layer, which is the reason why the requirements are considered from the early stages of the development process, thus making the development and maintenance easier, whilst fulfilling the project budget. The other methods, NDT [32] and Hera [38], meanwhile, have a development process that is based on Model-Driven Development (MDD). At this point it is important to highlight the work in [51], which presents a review (not an SLR) describing various MDWE (Model-Driven Web Engineering) methods (methods selected according to the author's own opinion without using selection criteria) that have been proposed, and discusses the advantages and disadvantages of these methods with regard to best practices in WS development. As our work focuses on RE activities and the tool support for development and the terminology used by each one, we did not extract information about the methods presented in the aforementioned review because it contained no relevant information that could be used to answer our research questions.

RQ2. - Which RE activities are supported by each method and the techniques used?

Almost all the methods analyzed in this SLR consider at least one of the RE activities, the exception being HERA since it does not have an explicit requirements phase. UWE, NDT and A-OOH are those methods which have placed greater importance on the RE activities by defining a set of formal guidelines to be used. The only method covering all the RE activities (elicitation, analysis, specification, validation and management) is NDT, which was initially created for the RE for hypertext applications and has been improved over the years to become a full WE method. The methods that do not place very much importance on RE activities are OOHDM and HERA since they only cover requirements specification. Requirements management is also one of the most important activities and one of those least covered, except by NDT [31] and A-OOH which support it by means of change impact analysis (CIA) [52], [43].

With regard to the techniques applied by each RE activity method, we have discovered that the methods use a specific set of technologies (Table 1) and that there would appear to be a trend toward the application of the UML (Unified Modeling Language) Use Cases, since OOWS, WebML, NDT, UWE and OOHDM use this technique in the requirements specification phase. There is also a trend toward the persistence of UML Profiles. The techniques extracted from the methods analyzed in this review are: Use Case Diagrams, a Data Dictionary, Conceptual Maps, the Functional Refinement Tree (FRT), UML-Profiles, Task Diagrams, Textual Templates, a Goal-oriented Requirements Engineering (GORE) modeling language named i*, Activity Diagrams, Interviews, Questionnaires and Checklists. Another technique that is widely accepted is that of Use Case Diagrams, which are used in traditional SE, and it is not surprising that WE methods also use them to model scenarios that may occur, when the user interacts with a WS. Moreover, with regard to UML, UML Profiles have recently

been adopted to provide a generic extension mechanism with which to customize UML models for particular domains, the technology used to do so being the Eclipse Modeling Project [53] by means of ECORE models. These profiles are applied by UWE [31], NDT and A-OOH [54]. Another UML technique that is applied by WebML and UWE is the Activity Diagram**,** which is a complementary technique for use case diagrams that is employed to model the logic captured by a single use case.

Table 1
The Requirements Engineering techniques used by each method

RE Technique	**UWE**	**NDT**	**WebML**	**OOWS**	**OOHDM**	**WSDM**	**A-OOH**
Use Cases	x	x	x	x	x		
Data Dictionary						x	
Conceptual Maps						x	
Functional Refinement Tree (FRT)				x			
UML-Profiles	x	x					x
Task Diagrams				x			
Textual Templates	x	x				x	
*i** Framework							x
Activity Diagrams	x	x	x				
Interviews	x	x					
Questionnaires	x						
Checklists	x	x					

GORE is applied in RE activities, and more specifically the *i** language [16] used by A-OOH [21] which is adapted by using an ECORE metamodel rather than UML-Profiles [54]. The *i** language has proved to be useful for representing: (i) the *stakeholders'* intentions, i.e., their motivations and goals, (ii) dependencies between *stakeholders* in order to achieve their goals, and (ii) the (positive or negative) effects of these goals on each other in order to be able to select alternative designs for the system, thus maximizing goal fulfillment. One of the basic problems with *i** is the growing requirements model (the scalability of the requirements model): when more requirements are set, the model tends to grow too much, and reading it therefore becomes complicated. This problem was recently solved [55] through the implementation of modules, i.e., the equivalent of UML Packages, in order to group the requirements according to two types of requirements (navigational and service). Another GORE technique used in WE is WebURN [57], a notation for early requirements analysis. NDT [27] and WSDM [58], meanwhile, apply Textual Templates for requirements specification. This technique is only applicable when the project is not very large; otherwise textual descriptions will grow significantly, thus making their maintenance and analysis

difficult. This technique can be applied in combination with use case diagrams, which is helpful for the developer depending on the level of granularity of the diagram in question.

The OOWS method uses Task Analysis and the Functional Refinement Tree (FRT) for requirements specification [59, 60]. Task Analysis is a hierarchical representation of which steps have to be performed in a task in order to achieve a goal. Since professionals often perform this, it usually depends on the analyst's experience. The FRT represents a hierarchical decomposition of the business functions of a system that is independent of the actual system structure. The authors of this method are currently working on a technique for the specification of requirements through the use of ontologies in order to solve this drawback.

The Conceptual Maps of Roles and Activities and a Data Dictionary [37, 58] techniques are used by WSDM. They are difficult to maintain and analyze owing to the fact that the requirements are basically defined in textual form. At this point, it is important to highlight the difference between a Data Dictionary and Textual Templates: a Data Dictionary is used to explain the semantics of words, the concepts or the terms that were used during the development process, while a Textual Template is a text structure defined by the methodology used in order to describe a particular functionality of the system that will be created, i.e., in use cases, a template can be added to describe the navigation of the WS, whereas the text descriptions, which are used in use cases, allow the artifact to be enriched, thus making it much easier to understand. Describing navigational requirements by means of textual descriptions is not an easy task owing to the description of alternative navigation paths through the Web system. The last technique is the User Interaction Diagram (UID), which is used by OOHDM to specify the interaction described in a Use Case for validation purposes and to support communication between the designer and users [28, 30, 61].

In summary, these techniques have advantages and disadvantages, e. g, the use of text for requirements specification in a complex development process is a disadvantage, because it is difficult to maintain, although it may nevertheless be extremely useful and comprehensible in the development of a simple Web system. With regard to Task Analysis, this is a set of techniques which is intended to provide a researcher with a complete understanding of what tasks a user really performs, what is needed to carry out those tasks, and what tasks a user should be doing, but this technique can be extremely time and resource consuming.

RQ3. - Is there any common terminology with regard to RE that is applied by the existing methods?

All disciplines needs a mutual terminology, which is required to allow researchers to understand and cooperate with each other, thus providing the basis for an improvement to the research and the reporting of processes in a particular research topic and making the findings from several empirical studies understandable. The research addressing requirements in WE has produced a heterogeneous

terminology for requirements that hinders further progress. A unified vocabulary, which is proposed in [7] (Section 2), is provided to shed light on (i) the expressivity of current methods when considering requirements in WE, and (ii) the correspondences between the custom terms applied to refer to requirements used by each method with the aforementioned classification. An overview of this is shown in Table 2, in which the cell labeled with an "X" indicates that the requirement is not explicitly considered by the method and the last column, labeled "NFRs", indicates that the method denominates non-functional requirements in general as "NFRs". In other words, the method does not use a specific term for each type of non-functional requirements. Of all the methods mentioned in the answer to RQ1, only A-OOH [21], UWE [56], WebML [45], OOWS [62], and NDT [27] cover all the types of requirements mentioned in [7]. However, they use their own terminology to denominate each type of requirement (its custom terms), with the exception of A-OOH, which applies the classification directly [54].

Table 2
Terminology used by each method in order to denominate its requirements

Classification / **Method**	**Content**	**Service**	**Navigational**	**Layout**	**Personalization**	**NFRs**
UWE	Content	Process	Navigation	Presentation	Adaptation	NFRs
NDT	Storage Information	Functional	Interaction	Interaction	Actor	NFRs
WebML	Content	Service	Navigational	Presentation	Personalization	NFRs
WSDM	Content	Functional	Navigational	X	Personalization	Security, Usability
OOWS	Functional	Functional	Navigational	Presentational	Presentation	NFRs
OOHDM	Content	X	Navigational	Layout	X	X
HERA	Content	Service	Navigational	Presentation	Personalization	X
A-OOH	Content	Service	Navigational	Layout	Personalization	Softgoal

Although the methods presented in Table 2 share, in a few cases, a term with the same name from the classification, some of them are used to consider an extra functionality, i.e., NDT includes Navigational and Layout requirements in the concept of Interaction Requirements and OOWS uses Content and Service requirements within FRs [59], [60], [62]. Finally, NFRs are considered in a very general way by basically all the methods. The only methods that consider NFRs in a more detailed form are WSDM [58] and A-OOH [63]: WSDM details Security and Usability NFRs and A-OOH considers the common types of NFRs within the

concept of "Softgoal", which is a general concept that can be used to represent any kind of NFRs in GORE.

RQ4. - Which methods provide tool support for their development processes?

All the methods provide tool support for their development process. NDT is supported by NDT-Suite [26, 27, 64], WSDM has WSDMtool [49, 65], WebML is supported by WebRatio [46, 66], UWE by MagicUWE [67], OOWS uses OlivaNova, OOHDM has OOHDM-WEB, Hera uses two tools, Saxon 7.0 and Sesame, and A-OOH has the WebREd-Tool [69]. With regard to RE activities, only NDT, UWE, OOWS and AOOH have tool support. NDT does this by means of NDT-Suite, UWE provides a Magic Draw plugin, the so-called MagicUWE, OOWS combines OOWS-Suite with the OlivaNova tool (deprecated) and OOWS-Suite [50], and A-OOH is supported by a set of Eclipse plugins [68], WebREd-Tool [40, 69], which won the best software demo award at the ICWE conference [40]. In terms of tool support for specific RE activities, the NDT, OOWS and A-OOH methods implement special techniques, i.e., A-OOH provides traceability support by means of an Eclipse plugin with which the requirements are specified, and when the generation of the conceptual models is performed, a weaving model is created to store the traces among requirements and the conceptual models, which is CIM-PIM in an MDA process [41]. NDT does this by means of the NDT-Suite, using traceability matrices [64]. OOWS uses two tools, the first of which is the open source tool called AGG (Attributed Graph Grammar System) and the second of which is called TaskTracer and is used to generate traceability reports [33]. With regard to the impact of changing a requirement, A-OOH supports CIA [52], which consists of verifying the impact resulting from the change made to any conceptual model after a requirements modification, while the WebREd-tool generates a report containing the requirements affected as the result of a change.

The answers to the RQs have allowed us to establish an analysis accompanied by suggestions for the dissemination of the methods studied in this SLR in the academic and industrial area, along with a list of suggestions for future research.

4.1 Dissemination of the Web Engineering Methods

The dissemination of the different methods is a highly important issue since it assists in the realization of important advances in the standardization of Web system development. The methods must be well known in both the academic world and the software industry. In this respect, it is worth mentioning the support offered by NDT, WSDM, UWE and WebML through their websites because all of them provide everyone who visits their websites with examples, published papers and their respective tools, with the exception of WSDM, which only offers the downloading of published papers because the tool's license is not free. In the particular cases of NDT, UWE and WebML, they provide guided step-by-step

examples with which to study and practice the development of WS using their respective support tools. This confirms why these methods are those most frequently used in academic and industrial projects. It is important to highlight the progress of the WebRatio tool (from WebML) since this has become an international spin-off enterprise and is derived from the definition of the IFML (Interaction Flow Modeling Language), which is the first language designed to express the content, user interaction and control behavior of the front-end of software applications, a standard designed by the OMG (Object Management Group) [70]. Several works have been carried out using this method, such as those presented in [71] and [72]. For more information about WebRatio, the authors' presents some lessons learned in [73].

4.2 Suggestions for Future Research

Empirical studies such as that by Nuseibeh [12] demonstrate that efforts made to provide a detailed business model with which to capture the *stakeholders'* requirements considerably reduce drawbacks in later phases of the development process. This idea will be used as a basis to conclude this analysis with some suggestions for future research: several of the WE methods do not support more than one RE activity, and this deficiency must be resolved, since the main purpose of RE is to facilitate the understanding of the product under development and the ability to manage any changes that occur during the development process. The development of robust tools for MDD covering the MDA life cycle is necessary, since MDD is a current trend (whose advantages have been studied [76]) and most WE methods implement it. The dissemination of the methods is a highly important issue in the standardization of WE. The normalization of the way in which requirements are denominated by each method is therefore necessary, because all disciplines need a mutual terminology, which is required to allow researchers to understand and cooperate with each other, thus providing the basis needed to improve the research area. Finally, although various studies regarding the benefits of MDD in a development process have been conducted [76], only a few of them refer to the WE domain ([51], [74] and [75]), which is why it is necessary to conduct more studies in order to validate and support its potential application.

Conclusions

This work presents the results obtained after carrying out an SLR. The aim was to create a comprehensive review and synthesis of the current state-of-the-art in the literature related to the RE activities in WE. To do this, a total of 3059 papers published in the literature and extracted from the most relevant scientific sources were considered, of which 14 were eventually analyzed in depth. The results of this SLR have shown that WE methods have not been designed to properly address development through the application of RE activities. What is more, they simply place less relevance on it or their corresponding techniques are poorly applied.

The relevance of a detailed and precise specification of requirements is well known; it helps to achieve an agreement with the customer, as regards to software functionality, user friendliness and priorities in the development process. However, the modeling of requirements does not occur in many projects, particularly in the Web domain, mainly owing to its special characteristics, its multidisciplinary development teams and its short time-to-market. Web systems can no longer be considered as common software systems owing to the diversity of users who access these applications. It is therefore necessary to provide solutions within the WE field, in order to specify and develop this type of system by considering the huge heterogeneous user population, their needs and goals. Our future work will include a MDD method guided by the RE activities, into which, techniques will be integrated for the automated generation of Web systems by means of an open source tool.

Acknowledgements

This work has been partially supported by the *Programa de Fomento y Apoyo a Proyectos de Investigación* (PROFAPI) from the Universidad Autónoma de Sinaloa (México), and the MANTRA project (GRE09-17) from the University of Alicante, Spain, and GV/2011/035 from the Valencia Government.

References

[1] Glinz, M.: 'On Non-Functional Requirements'. Proc. 15th Int. Req. Eng., Conf., Delhi, Oct. 2007, pp. 21-26

[2] Ginige, A., Murugesan, S.: 'Web Engineering: A Methodology for Developing Scalable, Maintainable Web Applications'. Cutter IT Journal, 2001, 14(7), pp. 24-35

[3] Cachero, C., Gómez, J.: 'Advanced Conceptual Modeling of Web Applications: Embedding Operation'. Proc. Proc. Jornadas de Ingeniería de Software y Base de Datos (JISBD), El Escorial, Spain, November 2002, pp. 235-248

[4] Casteleyn, S., Garrigós, I., Troyer, O. D.: 'Automatic Runtime Validation and Correction of the Navigational Design of Web Sites'. In Brown, S. (Ed.): 'APWeb' (IEEE Press, 2005, 7thedn.), pp. 453-463

[5] Ceri, S., Fraternali, P., Bongio, A.: 'Web Modeling Language (WebML): a Modeling Language for Designing Web Sites'. Int. Journal of Computer and Telecommunications Networking. 2000, 33(1-6), pp. 137-157

[6] Koch, N.: 'The Expressive Power of UML-based Web Engineering'. Proc. Int. Workshop on Web-oriented Software Technology (IWWOST), 2002, pp. 40-41

[7] Escalona, M. J., Koch, N.: 'Requirements Engineering for Web Applications - A Comparative Study'. Journal of Web Engineering, 2004, 2(3), pp.193-212

[8] Kitchenham, B.: ‘Procedures for Performing Systematic Reviews’ (Tech. rep., Keele University and NICTA, 2004)

[9] Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M, Bailey, J., Linkman, S.: ‘Systematic Literature Reviews in Software Engineering - a Systematic Literature Review’, Information and Software Technology, 2009, 51(1), pp. 7-15

[10] Aguilar, J. A., Garrigós, I., Mazón, J. N., Trujillo J.: ‘Web Engineering Approaches for Requirement Analysis- A Systematic Literature Review’. Proc. Web Information Systems and Technologies (WEBIST), Valencia, Spain, 2010, pp. 187-190

[11] Hull, E., Jackson, K., Dick, K.: ‘Requirements Engineering’. (Springer-Verlag, 2010)

[12] Nuseibeh, B., Easterbrook, S. M.: ‘Requirements Engineering: a Roadmap’. Proc. Int. Conf. on Software Engineering (ICSE), New York, 2000, pp. 35-46

[13] Maiden, N., Rugg, G.: 'ACRE: Selecting Methods for Requirements Acquisition’, Software Engineering Journal, 1996, 11, (3), pp. 183-192

[14] Shaw, M., Gaines, B.: ‘Requirements Acquisition’, Software Engineering Journal, 1996, 11, (3), pp. 149-165

[15] Viller, S., Somerville, I.: ‘Social Analysis in the Requirements Engineering Process: from Ethnography to Method’. Proc. IEEE Int. Symposium on Requirements Engineering, Limerick, June 1999, pp. 6-13

[16] Yu, E.: ‘Modelling Strategic Relationships for Process Reengineering’. PhD thesis, University of Toronto, 1995

[17] Bass, L., Merson, P., Clements, P., Bergey, J., Ozkaya, I., Sangwan, R.:, ´A Comparison of Requirements Specification Methods from a Software Architecture Perspective´. Software Engineering Institute, Carnegie Mellon University, 2006

[18] Gotel, O., Finkelstein, A.: ‘An Analysis of the Requirements Traceability Problem’. Proc. 1st International Conf. on Req. Eng., Colorado Springs, CO, Apr 1994, pp. 94-101

[19] S. E. Institute: ‘CMMI for Development: Guidelines for Process Integration and Product Improvement’, (Addison-Wesley Professional, 2011, 3rdedn.)

[20] Estublier, J.: ‘Software Configuration Management: a Roadmap'. Proc. Conf. on The Future of Software Engineering (ICSE), ACM, New York, NY, USA, 2000, pp. 279-289

[21] Garrigós, I., Mazón, J. N., Trujillo, J.: ‘A Requirement Analysis Approach for Using i* in Web Engineering’. Proc. Int. Conf. on Web Engineering (ICWE), San Sebastian, Spain, June 2009, pp. 151-165

[22] Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., Khalil, M.: 'Lessons from Applying the Systematic Literature Review Process within the Software Engineering Domain'. Journal of Systems and Software, 2000, 80, pp. 571-583

[23] Walia, G. S., Carver, J. C.: 'A Systematic Literature Review to Identify and Classify Software Requirement Errors', Information and Software Technology, 2009, 51, (7), pp. 1087-1109

[24] Beecham, S., Baddoo, N., Hall, T., Robinson, H., Sharp, H.: 'Motivation in Software Engineering: A Systematic Literature Review'. Information and Software Technology. August 2008, 50, (9-10), pp. 860-878

[25] Fons, J., Valderas, P., Ruiz, M., Rojas, G., Pastor, O.: 'OOWS: A Method to Develop Web Applications from Web-oriented Conceptual Models'. Proc. Int. Workshop on Web Oriented Software Technology (IWWOST), sprint, 2003, pp. 65-70

[26] Escalona, M. J, Mejia, M., Torres, J.: 'Developing Systems with NDT & NDT-Tool'. Proc. Int. Conf. on Information Systems Development: methods and tools, theory and practice, Vilna, Lithuania, 2004, pp. 149-59

[27] Escalona, M. J., Aragón, G.: 'NDT. A Model-Driven Approach for Web Requirements'. Trans. on Software Engineering, 2008, 34(3), pp. 377-390

[28] Garzotto, F., Paolini, P., Schwabe, D.: 'HDM–a Model-based Approach to Hypertext Application Design'. ACM Transactions on Information Systems (TOIS), 1993, 11(1), 1-26

[29] Schwabe, D., de Almeida Pontes, R., Moura, I.: 'OOHDM-Web: an Environment for Implementation of Hypermedia Applications in the WWW', ACM SIGWEB Newsletter, 1999, 8, (2), pp. 18-34

[30] Schwabe, D., Rossi, G.: 'The Object-Oriented Hypermedia Design Model', Communications of the ACM, 1995, 38, (8), pp. 45-46

[31] Escalona, M. J., Koch, N.: Metamodeling the Requirements of Web Systems. In: Int. Conf. on Web Information Systems and Technologies (WEBIST), Lecture Notes in Business Information Processing, Vol. 1, July 2007, pp. 267-280, Springer Berlin Heidelberg

[32] Koch, N., Zhang, G., Escalona, M. J.: 'Model Transformations from Requirements to Web System Design'. Proc. Int. Conf. on Web Eng. (ICWE), ACM, New York, NY, USA, 2006, pp. 281-288

[33] Valderas, P., Pelechano, V.: 'Introducing Requirements Traceability Support in Model-driven Development of Web Applications', Information and Software Technology, 2009, 51, (4), pp. 749-768

[34] Molina, F., Toval, A.: 'Integrating Usability Requirements that can be Evaluated in Design Time into Model-driven Engineering of Web

Information Systems', Advances in Engineering Software, 2009, 40, (12), pp. 1306-1317

[35] Bolchini, D., Mylopoulos, J.: 'From Task-oriented to Goal-oriented Web Requirements Analysis'. Proc. Int. Conf. on Web Information Systems Engineering (WISE), IEEE Computer Society, Washington, DC, USA, Dec 2003, pp. 166-175

[36] Koch, N.: 'Transformation Techniques in the Model-driven Development Process of UWE'. Proc. Workshop of the Int. Conf. on Web Eng. (ICWE), ACM, New York, NY, USA, 2006

[37] De Troyer, O. M. F., Leune, C. J.: 'WSDM: a User-centered Design Method for Web Sites'. Int. Journal of Computer and Telecommunications Networking, 1998, 30(1-7), pp. 85-94

[38] Houben, G., Barna, P., Frasincar, F., Vdovjak, R.: 'Hera: Development of Semantic Web Information Systems'. Proc. Web Eng., Oviedo, Spain, July 2003, pp. 529-538

[39] Gómez, J., Cachero, C., Pastor, O.: 'Extending a Conceptual Modeling Approach to Web Application Design'. Proc. 12th Int. Conf. on Advanced Information Systems Eng. (CAiSE), Springer-Verlag, London, UK, June 2000, pp. 79-93

[40] Aguilar. J. A., Garrigós, I., Casteleyn, S., Mazón, J. N.: 'WebREd: A Model-driven Tool for Web Requirements Specification and Optimization'. Proc. Int. Conf. on Web Engineering (ICWE), Lecture Notes in Computer Science, Springer Berlin Heidelberg, May 2002, pp. 452-455

[41] Aguilar, J. A., Garrigós, I., Mazón, J. N.: ´Modelos de weaving para trazabilidad de requisitos Web en A-OOH´. Proc. DSDM: Actas del VII Taller sobre Desarrollo de Software Dirigido por Modelos, JISBD, Congreso Español de Informática (CEDI), SISTEDES, Valencia, España, 2010, pp. 146-155

[42] Aguilar, J. A., Garrigós, I., Mazón J. N., Trujillo, J.: 'An MDA Approach for Goal-oriented Requirement Analysis in Web Engineering'. Journal of Universal Computer Science (JUCS), 16, (17), pp. 2475-2494

[43] Aguilar, J. A., Garrigós, I., Mazón, J. N., Zaldívar, A.: 'Dealing with Dependencies among Functional and Non-functional Requirements for Impact Analysis in Web Engineering'. Proc. Int. Conf. on Computational Science and Its Applications (ICCSA), Springer-Verlag, Salvador de Bahía, Brazil, June 2012, pp. 116-131

[44] Casteleyn, S., Van Woensel, W., Houben, G. J.: 'A Semantics-based Aspect-oriented Approach to Adaptation in Web Engineering'. Proc. Conf. on Hypertext and Hypermedia (HT), ACM, New York, NY, USA, 2007, pp. 189-198

[45] Bongio, A., Milano, P. D., Fraternali, P., Maurino, A., Ceri, S.: ´Modeling data entry and operations in WebML´. Proc. World Wide Web and Databases (WebDB), TX, USA, May 2000, pp. 201-214

[46] Brambilla, M., Butti, S., Fraternal, i P.: 'WebRatio BPM: A Tool for Designing and Deploying Business Processes on the Web'. Proc. Int. Conf. on Web Engineering (ICWE), Lecture Notes in Computer Science, Springer Berlin / Heidelberg, July 2010, pp. 415-429

[47] Brambilla, M., Fraternali, P.: ´Implementing the Semantics of BPMN through Model-driven Web Application Generation´. Proc. Business Process Model and Notation, Lucerne, Switzerland, Nov 2011, pp. 124-129

[48] Vdovjak, R., Frasincar, F., Houben, G., Barna, P.: 'Engineering Semantic Web Information Systems in Hera', Journal of Web Engineering, 2003, 2, pp. 3-26

[49] Quintero, R., Pelechano, V., Pastor, O., Fons, J.,: 'Aplicación de MDA al Desarrollo de Aplicaciones Web en OOWS'. Proc. Jornadas de Ingeniería de Software y Base de Datos (JISBD), 2003, pp. 84-668

[50] Valverde, F., Valderas, P., Fons, J.: 'OOWS Suite: Un Entorno de desarrollo para Aplicaciones Web basado en MDA'. Proc. Workshop Iberoamericano de Ingeniera de Requisitos y Ambientes Software (IDEAS), Isla Margarita, Venezuela, 2007

[51] Hincapié, J. A., Duitama F.: 'Model-driven Web Engineering Methods: a Literature Review'. Rev. Fac. Ing. Univ. Antioquia, 2012, 63(1), pp. 69-81

[52] Aguilar, J. A., Garrigós, I., Mazón, J. N.: 'Impact Analysis of Goal-oriented Requirements in Web Engineering'. Proc. Int. Conf. on Computational Science and Its Applications (ICCSA), Springer-Verlag, Santander, Spain, June 2011, pp. 421-436

[53] 'Eclipse Modeling Project', http://www.eclipse.org/emf, 2014

[54] Aguilar, J. A.:'A Goal-oriented Approach for Managing Requirements in the Development of Web Applications'. PhD Thesis. University of Alicante, Spain, 2012

[55] Aguilar, J. A., Zaldívar, A., Tripp, C., Misra, S., Sánchez, S., Martínez, M., García, O.: 'A Solution Proposal for Complex Web Application Modeling with the I-Star Framework'. Proc. Int. Workshop on Software Engineering Process and Applications, Porto, Portugal, July 2014, pp. 135-145

[56] Koch N., Kozuruba, S.: 'Requirements Models as First Class Entities in Model-driven Web Engineering'. Current Trends in Web Engineering, Berlin, Germany, July 2012, pp. 158-169

[57] Chawla, S., Srivastava, S.: 'A Goal-based Methodology for Web Specific Requirements Engineering'. Proc. of World Congress on Information and Communication Technologies, Trivandrum, Oct-Nov 2012, pp. 173-178

[58] Troyer, O. D., Casteleyn, S.: 'The Conference Review System with WSDM'. Proc. Int. Workshop on Web oriented Software Technology (IWWOST), 2001, pp. 30-98

[59] Durán, G. E. R.: 'Modeling Adaptative Web Applications in OOWS'. PhD thesis, Technical University of Valencia, 2008

[60] Abrahao, S., Fon, J., González, M., Pastor, O.: 'Conceptual Modeling of Personalized Web Applications', in Springer Berlin / Heidelberg, Proc. (Ed.): 'Adaptive Hypermedia and Adaptive Web-based Systems' (IEE Press, 2002, 2nd), pp. 358-362

[61] Martín, A., Rossi, G., Cechich, A., Gordillo, S.: 'Engineering Accessible Web Applications', An Aspect-oriented Approach. World Wide Web, 2010, 13, (4), pp. 419-440

[62] Fons, J., Garca, F., Pelechano, V., Pastor, O.: 'User Profiling Capabilities in OOWS'. Proc. Int. Conf. on Web Engineering (ICWE), Springer Berlin Heidelberg, July 2003, pp. 486-496

[63] Aguilar, J. A., Misra, S., Zaldívar, A., Bernal, R.: 'Improving Requirements Specification in WebREd-Tool by Using a NFR's Classification'. Proc. Int. Workshop on Software Engineering Process and Applications, Salvador de Bahía, Brazil, July 2013, pp. 59-69

[64] Garca-Garca, J., Alba Ortega, M., Garca-Borgoon, L., Escalona, M.: NDT-suite.: 'A Model-based Suite for the Application of NDT. Proc. Web Engineering, Germany, Berlin, July 2012, pp. 469-472

[65] Wilder, K. V.: 'Implementation Generation for WSDM using Web Applications Framework'. PhD thesis, University of Brussels, 2009

[66] Acerbis R, Bongio A, Brambilla M, Butti S, Ceri S, Fraternali P.: 'Web Applications Design and Development with WebML and WebRatio 5.0'. Proc. Objects, Components, Models and Patterns, Springer Berlin Heidelberg, June-July 2008, pp. 392-411

[67] Busch, M., Koch, N.: MagicUWE.: 'A CASE Tool Plugin for Modeling Web Applications'. Proc. Int. Conf. on Web Engineering (ICWE), Springer-Verlag, Berlin, Heidelberg, June 2009, pp. 505-508

[68] 'Eclipse', http://www.eclipse.org/, 2015

[69] 'WebREd-Tool', http://webred.maz.uasnet.mx/, 2015

[70] 'Interaction Flow Modeling Language', http://www.ifml.org/, 2015

[71] Ingle, D., Meshram, B.: 'Analyzing Web Modeling Existing Languages and Approaches to Model Web Application Design'. International Journal of Advanced Research in Computer Engineering & Technology, 2012, 1(3), pp. 196-204

[72] Silvera, J. A., Arias, D., Gil, G.:'Ingeniería de software aplicada a un sistema de gestión de calidad en centros educativos'. XV Workshop de Investigadores en Ciencias de la Computación, La Plata, Argentina, May 2013, pp. 517-520

[73] Brambilia, M., Fraternali, P.: 'Large-Scale Model-driven Engineering of Web User Interaction: the WebML and WebRatio Experience'. Science of Computer Programming, 2014, 89(B), pp. 71-87

[74] Valderas, P., Pelechano, V.: 'A Survey of Requirements Specification in Model-driven Development of Web Applications'. ACM Transactions on the Web (TWEB), 2011, 5, (2)

[75] Méndez, E.: 'A Systematic Review of Web Engineering Research'. Proc. Int. Symposium on Empirical Software Eng. (ESEM), New Zealand, Nov. 2005, p. 10

[76] Martínez, Y., Cachero, C., Meliá, S.:'MDD vs.traditional software development: a practitioners subjective perspective. Information and Software Technology, 2012, 55(2), pp. 189-200

Impact of Reproduction Size and Halftoning Method on Print Quality Perception

Ivan Pinćjer[1], Dragoljub Novaković[1], Uroš Nedeljković[1], Nemanja Kašiković[1], Gojko Vladić[1]

[1]University of Novi Sad, Faculty of Technical Sciences, Department of Graphic engineering and design, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
pintier@uns.ac.rs, novakd@uns.ac.rs, urosned@uns.ac.rs, knemanja@uns.ac.rs, vladicg@uns.ac.rs

Abstract: This paper presents a study related to image reproduction quality assessment. The experiment was designed to determine the impact of halftone image dimensions on the perception of image fidelity and grain structure, these two, being important quality attributes. Subjective quality assessment experiments were designed to complement recently published findings obtained by objective image metric methods for quality assessment. Image fidelity and grain structure are understandable to the observers that are not familiar with the methods used to determine image quality. These attributes are directly dependent on the chosen halftoning method. In this research, the samples were halftoned using two different types of screening methods: frequency modulation (FM) and amplitude modulation (AM) method. The experiment produced around 4000 data samples, which were analyzed by standard statistical methods. Results indicate a significant influence of the image size and halftoning method on the subjective quality assessment.

Keywords: halftoning; size of the reproduction; psychophysical experiment; print quality; image quality

1 Introduction

One of the indispensable elements of the graphic process, which always makes for an interesting topic of research, both for the scientists and industry, is halftoning. The objective of halftoning is to ensure consistent printing, along with the fidelity of reproduction. As the image consists of raster dots, it is clear that the image quality is directly dependent on the parameters that are related to the raster element. When the structure of the various IQ (image quality) metrics are analyzed, the influence of the rasterization on the image quality is somewhat disregarded. In order to use metrics for the comparison of printed image and the original, the reproduction needs to be digitalized. The goal of using metrics is to replace the human subjective factor in the evaluation of the print quality, as well

as, to speed up the process of the evaluation mentioned above. The way in which metrics attempts to accomplish this is through the removal of halftones from the image. Human Visual System (HVS) is implemented by taking the image through the process of spatial filtering based on contrast sensitivity function [1]. When using the HVS, print analysis with image metrics, largely depends on the chosen method of filtering. Image metrics that use HVS is therefore not suitable for investigation of the halftone structure.

In the area of scientific research, many differences between the conventional AM screen and stochastic FM screen have been perceived and objectively measured so far [2] [3]. Color gamut, the ability of reproduction of uniform tone and color gradients [4], contrast, sharpness, the visible pattern [5]-[8] are some of the most important analyzed attributes. The first goal of this research was to show to what extent the objectively measured differences are perceived by observers. The second goal was to determine if the perception is influenced by the size of the image, keeping the viewing distance and screen ruling constant and making only the viewing angle variable.

2 Formulation of the Research Problem and Hypothesis

AM halftones have a constant dot frequency while dot amplitude is dependent upon the image gray level. FM halftone dots have fixed amplitude, but the dot distribution is variable [9] [10]. FM algorithms provide a very good representation of the original contone image, considering discretization and loss of data. The result is a high frequency isolated dot. Continuous and repeatable printing of such dot is possible only with high-quality printing systems or digital ink-jet systems with a good addressability. In the offset printing technique, imaging of printing plates is controlled directly by the computer software. Modern computer to plate (CTP) high-resolution devices are up to the challenge of imaging minute isolated dot. Research and development of different rasterization techniques were driven by the need for practical application. Apart from making the printing process possible, the goal of the halftoning is to provide the best possible quality for the given conditions. Problem emerges in the process of ink transfer from the inking unit to the printing plate, then from the printing plate to the blanket cylinder, and finally, within ink transfer process from the blanket to the substrate. The solution was a reduction of frequency using white noise middle frequencies called “green noise” [11]. Green noise is characterized by its aperiodicity, lack of low-frequency grainy structure [12], and clustering of halftoning elements. Clustering of pixel increases the halftone element, which is precisely the key to the possibility of application of the stochastic rasterization in conventional offset printing, as well as on the devices with high unreliability. This kind of approach to the solution of

the problem is defined as AM/FM rasterization in research papers [13] or, more commonly, as a second generation FM. Application of the second generation FM halftone provides the possibility of defining a cluster size, i.e. its coarseness to suit various offset systems or even different printing technique like ink-jet digital printing. The final print quality in digital printing depends on: image processing, inks used, printing machine, substrate, and number of layers printed [14]. With unreliable devices, the coarseness can be increased in such a way to obtain larger halftone elements with greater constancy. The visibility of halftone can be reduced with high precision devices. By simple merging of AM halftone in mid-tones with FM halftone in highlights and shadows, a so-called hybrid halftone [2] was created. Hybrid halftone finds its application in various printing techniques, but due to its construction and specificity it will not be the object of this research. The study of image quality is multidimensional and multidisciplinary area. Print image quality can be measured objectively and subjectively. Objective print quality measuring methods are based on physical print measuring, i.e. use of measuring instruments (spectrophotometer, densitometer...), by which printing parameters, defined by external (ISO 12647, GRACol, PSO, Fogra, System Brunner...) or internal standards, are achieved. All of these standards, besides the measurement ranges, also contain photographs on test sheet for visual print evaluation. That shows the value of the subjective evaluation of the quality. Considering subjective evaluation of the quality is demanding, both in terms of time and human resources, a need emerged for development of an objective evaluation of the quality, that would replace the subjective one, but still be in high correlation with it. In order to achieve this, next to the mathematical calculation of error or noise on the image, it is also necessary to implement the influence of human visual system (HVS) [15]. The role of the HVS filter is to provide the information on whether a certain noise frequency will be visible to the human eye at a certain distance and to integrate its estimate in the evaluation of the quality of the reproduction, along with the measured values. However, contrast sensitivity function cannot be directly applied due to the complex structure of images, which is why the best algorithms for image quality evaluation do not contain HVS, but are instead based on structural or informational differences [16]. Based on the analysis so far, it can be asked if there are any other parameters that influence the perception of halftone structures with the observers. On the other hand, the size of the halftone dot, both with the AM and FM, is constructed in such a way as to not be visible to the human eye. The visibility threshold is set at four cycles per the degree of the visual angle [17]. This number of cycles correlates with the screen rulings of 110lpi. Modern offset printing processes can easily achieve screen rulings of 150 and 175lpi, which amounts to six and seven cycles, respectively, per the degree of the angle of perspective. This frequency is higher than the threshold frequency, so it can be assumed that the raster will not be visible, and therefore that the observer cannot perceive the difference in the quality of reproduction rasterized by different raster types (AM and FM) and their variations in frequency.

Based on the analysis of the available research and perceived problems in the process of reproduction of the original, two hypotheses were formulated that were tested during our research (H1 and H2) and which will be presented concisely:

H1: During the process of subjective evaluation, perception of the image quality attributes *grainy structure, sharpness and smoothness,* are influenced by halftoning method.

Differences in the properties of AM and FM halftoning can also be seen through the discretization of a contone image. The loss of information of the continuous tone image is the result of amplitude reduction to a one-bit information depth. The only way to preserve the data quantity is to increase the data frequency. With its high-frequency, FM halftone provides the preservation of a larger data quantity then AM, while at the same time preserving the information that contributes to greater image sharpness and smooth texture transitions. By changing the size of the reproduction, i.e. by increasing it, the quantity of lost information, due to halftoning, is reduced. It can be expected that with a certain increase of reproduction dimensions this advantage of FM halftone will be reduced. Dimension increase of image reproduction would consequently change the image quality perception of the observers. It can also be expected that the reproduction with higher screen ruling, of the same type of halftone, will obtain better results when evaluated by the observers, since it displays a larger quantity of information.

H2: During the process of subjective evaluation, perception of the image quality attributes *grainy structure, sharpness and smoothness* are influenced by the size of the reproduction.

Answers to research questions have to be resolved via subjective evaluation of the reproduction quality in an experiment with observers. According to the attributes defined in the epistemological study done by Pedersen and colleagues [18], grainy structure of the images, i.e. noise, sharpness and smoothness make up three of the five basic attributes of image quality, next to the color and lightness. The attributes of *sharpness*, *smoothness* and *noise* can serve as parameters for the evaluation of the quality of the reproduction done with different types of halftone, and they can be found in the evaluated parameters *the least noticeable grainy structure* and *the highest fidelity of reproduction*.

The hypotheses formulated in this paper were tested via observer's evaluations according to given parameters. The first parameter is tied to the graininess of the image [19]. The graininess of the image is defined as the noticeable low-frequency pattern of a periodic or stochastic structure. According to the literature, it is defined as a negative influence on the overall picture quality [20]. During the halftoning, i.e. discretization of the continuous tone image, undesirable patterns unavoidably occur [21], which is why their reduced noticeability is defined as an attribute of quality. Test category of the *least noticeable grainy structure* examines the quality of halftoning concerning this attribute.

The second parameter related to the evaluation of the reproduction is fidelity [22]. The goal of the examination of the high fidelity structure is to choose the image that shows the least negative attributes of halftoning.

Apart from the method of rasterization, there are other physical parameters that influence the perceived print quality, as the type of the substrate, printing technique, color, etc. [23]. In the research, reproduced images halftoned both by AM and FM screen were evaluated, with preservation of exactly the same values of parameters that can influence print quality. During the experiment, attention was directed towards a single variable, by keeping the other parameters constant. In this manner, the possible differences in visual experience of the observers, in the course of subjective evaluation, will focus exclusively on the attribute that is being varied, i.e. size of the reproduction. Test reproductions were generated in a strictly controlled laboratory environment, in order to exercise the greatest possible amount of control over all of the parameters. Psychophysical testing was conducted under standard experiment conditions while the results were processed with accepted statistical methods for independent categorical variables.

3 The Goal and Purpose of the Research

Changes in the viewing distance and the influence of the distance on the HVS were examined in order to integrate HVS in some of the accepted metrics for the evaluation of the image quality [24]. The goal of this research was to point to a single important parameter that influences the evaluation of the quality of the reproduction. The research also has a direct application in real printing systems, when considering guidelines in choosing the appropriate halftoning technique.

The purpose of the research is to develop a better understanding of the parameters that influence image quality metrics for printed images. For the purposes of this research, a unique experiment was constructed, which was carried out with the observers, both with and without their knowledge of the reproduction process. Observers type entailed the development of a test that would be understandable both to the professional and the naïve observer. The results are elaborated in a previous study by comparing the two observers group. It can be concluded that the test was appropriately constructed since there was no statistically significant difference between the answers provided by professional and naïve observers [7].

This research can contribute to the discussion about cost justification of implementing the new halftoning techniques, with more detailed view of the cost validity in relation to the production needs.

4 Description of Experimental Framework

The research was conducted utilizing a subjective quality assessment experiment. The observers had to evaluate the quality of the reproductions subjectively by using test samples with reproductions printed in four different halftones and three different sizes. In the course of choosing the halftoning method, second-generation FM halftones were considered [8]. The algorithms used were created outside the conventional CTP systems. Image halftoning of testing samples was conducted by software Raster Image Processor, compatible with PostScript 3 and PDF 1.7 files, and based on GhostScript PostScript/PDF engine. This procedure for screening was chosen because of its flexibility to obtain necessary results. For the AM raster, the Euclidian dot was used, which is the most frequently used shape in everyday printing.

The fineness of amplitude-modulated halftone is described by the number of lines per inch (lpi) while the FM halftone is defined by the size of the microdots (μm). The screen ruling and sizes of the FM raster dot that were used in the experiment were chosen based on the previous research [6]. Differences perceived by the observers will be the result of the size variable exclusively. According to the given instructions, every observer looked at thirty-six different reproductions and ranked them [25] in relation to the two parameters: *the least noticeable grainy structure* and *fidelity of the reproduction.* After the viewing process, observers were instructed to fill in the questionnaire.

The experiment was conducted with twelve different test sheets: one test was made for each of the four images distinguishable by the content in each of the three sizes. The sizes of the images were chosen in such a way as to simulate reproduction on a packaging or in a magazine. Three reproduction sizes were chosen: 62 × 44 mm, 88 × 62 mm and 125 × 88 mm, consequently creating three different viewing angles. Therefore, on each of the three tests sheets with different reproduction sizes there are four pictures of the same content, same size, but different halftoning method.

In order for the research to proceed, four test sheets were created, where on each of the test sheet an image of a different iconic content was reproduced. One image was presented on one test sheet but reproduced with different halftoning algorithms: two conventional (150lpi and 175lpi) and two stochastic (20 μm and 40 μm). All four test sheets with their belonging images were reproduced in three different sizes. Observers then used the ranking method and graded the observed samples, based on the test questions that they were given. The reproductions were prepared in such a way as to expect certain preferences from the observers towards a specific type of halftone. The shift in preferences on different test sheets is supposed to discover if the change in the size of the reproduction has an influence on the change in the preferred type of raster.

The transfer of the images from a digital to an analog form, had to conform to the same principles of the reduction of influential parameters to the minimum. Reduction of influential parameters entailed that all images, no matter what type of halftone was used, had to be reproduced in the same manner and under the same conditions and using the same substrates. The solution demanded the use of a printing device that would be able to reproduce the desired halftone, without modifying halftone parameters. Software RIP (Raster Image Processor) would have to override device RIP to preserve defined parameters. The reproductions were printed on a machine for print proof Epson Stylus Pro 7800, adhering to demands for the experiment. The resolution of the output device was set to 1.440 dpi, which is the maximum resolution of this printing device and enables the printing of the needed sizes of the halftone dots. Color profiles used were ISO Fogra 39 for AM and ISO Fogra 43 for stochastic one.

The printing of the samples, to be used in the test, was preceded by the calibration of the printing proof device, in order to simulate offset printing. The calibration of the output device was done with RIP software, which was to be used for the reproduction of the halftoned images. After the linearization, a halftone proofing was done. This kind of control was necessary to ensure that the printed halftone will match its digital equivalent exactly and for each dot, i.e. dot-by-dot. Comparison of the printed halftone with its digital variant was done for every type of halftone separately.

The last step in the print standardization was the calibration of the output device. Calibration was done using TECHKON SpectroDens, and, as a result, calibration curves were obtained for every color and every type of rasterization used in the experiment. The printed samples were measured by the measurement device Techon Spectroplate to confirm the screen ruling and structure of halftone dots of the printed reproductions with given parameters.

The pictures were chosen in such a way as to be mutually exclusive in the frequency of detail, smoothness, and their combination, in line with experiments of the authors that have done distinguished work in the objective analysis field of the reproduction quality [20], [26], and [27]. The four images presented on *Figure 1* were chosen.

Figure 1
Images used in the experiment: a girl (test reproduction 1), coffee (test reproduction 2), a plate (test reproduction 3), a car (test reproduction 4)

The chosen images are appropriate for the evaluation of the examined attributes of quality, a girl – close-up of a human face, coffee – large number of details,

background is out of focus, a plate – many details, uniform colors, a car – presentation of a front of the car, very little detail, uniform smoothness.

The printed reproductions were placed on a neutral gray material and presented to the observers. The observers looked at the 12 panels with the reproductions, one by one and marked images, as instructed at the beginning of the experiment. In order to standardize conditions, the panels were observed in a viewing cabin under the controlled light. The strength of the light was 2200 lx, and the color temperature was 5000 K. The viewing distance was set at 40 cm. Observers analyzed four images on a single test sheet, at the same time, and subjectively evaluates the image that best reflects the given criteria: *grainy structure* and *fidelity of reproduction*. At the moment of observation, the examinee has no information about the difference between the observed reproductions, nor the type of the halftone used to produce each of the reproductions. Every test was prepared with a different allocation of images so that one type of halftone changes its position in the test. Changes in layout eliminate the possibility for the observer to exploit a pattern. The observers had unlimited time to look at the reproductions, and, respectively, for their evaluation in line with the test questions. There were a 101 observers, out of which 50 were female and 51 male, with the average age of 24.7 years. Sixty-eight of them having normal vision and 33 having their vision corrected to normal.

Since the test compares independent attributive non-parametric data, the results were processed with the Chi-square statistical analysis using SPSS Statistics v.20. These are quantitative data, i.e. frequency of choosing the observed parameter of the quality of the image for a specific independent group, respectively halftone type. This kind of statistical processing will enable determination of a significant statistical difference between the perception of the same halftone type on different sizes of the test images, as well as different halftone types on the same sample sizes. Furthermore, if the null hypothesis is accepted the dimensions of the test samples, i.e. images would have no effect on the evaluation of the parameters given. Consequently, the distribution of the observer`s answers would be the same for all three sizes, whether the distribution is coincidental or dependent upon the halftoning of the reproduction. The frequency of answers on the given parameters and their distribution according to different halftone types (20 μm FM, 40 μm FM, 150lpi AM and 175lpi AM (Fig. 2)) can also be followed.

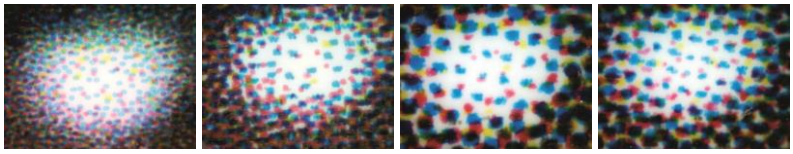


Figure 2

Magnified detail (140 ×) of the printed test reproduction 1 prepared with different halftone methods (20 μm FM, 40 μm FM, 150lpi AM and 175lpi AM)

5 Results and Discussion

In order to see which of the given images the observers evaluate as having the least noticeable grainy structure, and which ones they evaluate as having the highest fidelity of reproduction, a descriptive statistic was done for every picture separately.

A Chi-square test was calculated for each of the test sheets. A significant interaction was found for statistical significance $p<.05$. There were 24 Chi-square values in total, 12 for the least noticeable grainy structure and 12 for the highest fidelity reproduction. Table 1 and 2 shows results for test sheet 1.1 and 1.2.

Table 1
The least noticeable grainy structure for test reproduction 1.1

	Frequency	Percent
A (20 μm)	85	84.1
B (40 μm)	3	3.0
C (150lpi)	2	2.0
D (175lpi)	9	8.9
No difference	2	2.0
Total	101	100.0

Table 2
The least noticeable grainy structure for test reproduction 1.2

	Frequency	Percent
A (20 μm)	81	80.1
B (40 μm)	3	3.0
C (150lpi)	5	5.0
D (175lpi)	11	10.9
No difference	1	1.0
Total	101	100.0

After the analysis of each test sheet separately, a cross tabulation was done. Cross tabulation gives insight into the answer distribution for each of the reproduction sizes, so it facilitates the tracking of change in the choice depending on the change in reproduction size. The Chi-square was calculated for the reproductions of different sizes by using a contingency table.

Data analysis by a Chi-square statistical method allowed the rejection of null hypothesis and acceptance of H1 hypothesis: There is a statistically significant difference in the perception of the quality attributes of *the least noticeable grainy structure* and *the fidelity of reproduction,* dependent on the halftoning method of the image, with 95% certainty. With all of the test sheets, a statistically significant difference in the choice of preferred reproduction is shown, supporting the notion that the observers had discerned the varied parameters of the rasterization type. Finer AM screen ruling, as well as higher FM frequencies, significantly stand out in a choice of the same types of halftoning with course screen and lower frequencies. Implementation of a graphic system that could support fine screen ruling or frequency, would significantly improve the print quality. The second hypothesis analyzes the change of perception according to the two given parameters of *the least noticeable grainy structure* and *the fidelity of reproduction* depending on the image size. Each image content is analyzed separately.

Table 3
Comparative analysis of different sizes of the test reproduction 1 for the parameter *the least noticeable grainy structure*

	A (20 µm)%	B (40 µm)%	C (150lpi)%	D (175lpi)%	No difference%
1.1	84.1	3	2	8.9	2
1.2	80.1	3	5	10.9	1
1.3	7.9	2	22.8	66.3	1

χ^2=165.271, df=8, *p*-value: 0.00

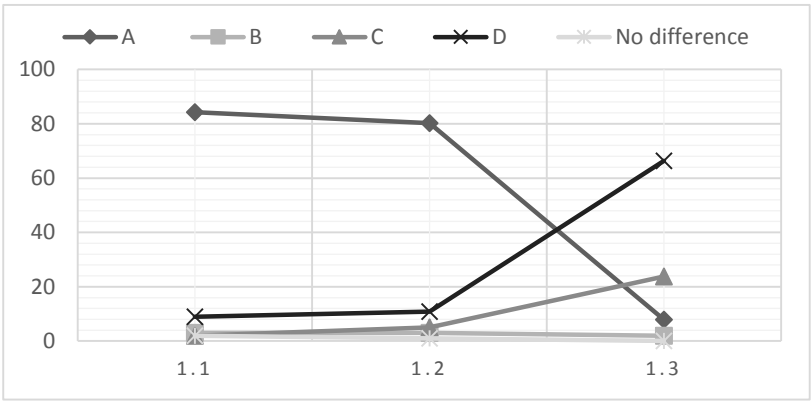


Figure 3
Observer's answers according to the parameter *the least noticeable grainy structure* for different sizes of the test reproduction 1

Table 3 shows that the observers mostly chose the FM raster (84.2%). As far as 83.5% of the observers chose 20 µm FM for the first two sizes, whereas 71.8% of observers, who chose FM raster with the smallest image size, decided upon 175lpi AM raster with the biggest image size. Therefore, the percentage of the observers who chose 175lpi AM raster with the biggest size went up to 66.3%. It can be concluded that the smaller the image is, the bigger the need is to use FM raster, by analyzing the data obtained from the test reproduction 1 (a girl). However, the situation changes drastically when the observers are shown the same image, with the same halftoning, on the same substrate, observation conditions, but with different size (Fig. 3).

Analyzing the results of 1.1 and 1.2 with a Chi-square, it was determined that there was no statistically significant difference between two images. However, the difference emerged between the middle (1.2) and the largest size (1.3). Image size 125 × 88mm has enough information to display the image with the 175lpi AM without too much data loss. As soon as the image became large enough, as to contain all the important detail, the negative attribute of FM manifested as image

noise, asserted itself, while the uniform AM structure became dominant when speaking about the least noticeable grainy structure on the largest image.

Table 4
Comparative analysis of different dimensions of the test reproduction 2
for the parameter *the least noticeable grainy structure*

	A (20 μm)%	B (40 μm)%	C (150lpi)%	D (175lpi)%	No difference%
2.1	60.4	5	9.8	10.9	13.9
2.2	71.2	5	5	15.8	3
2.3	78.3	5.9	6.9	3	5.9

χ^2=21.217, df=8, *p*-value: 0.00659279

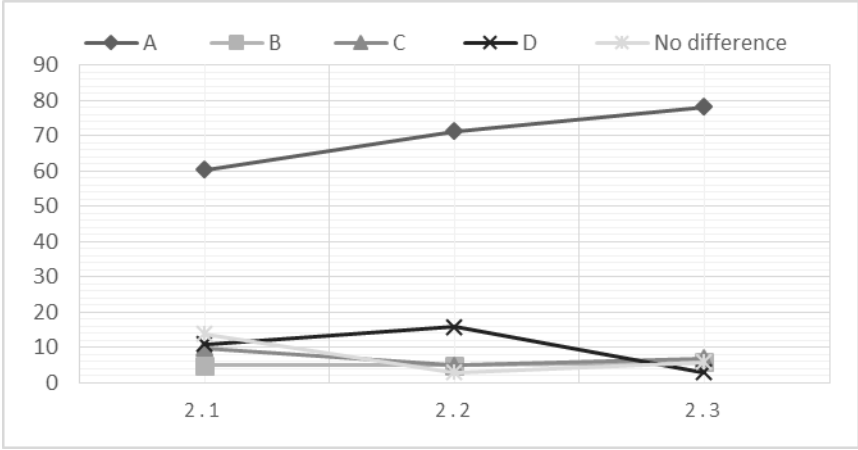


Figure 4
Observer's answers according to the parameter the *least noticeable grainy structure* for the different dimensions of the test reproduction 2

Table 4 shows the low effect of different image size on image content with high frequencies. In the case of the smallest image, most of the observers chose the 20 μm FM halftone at 60.4%, whereas this number went up to 78.3% for the largest test reproduction size. Values for the other types of halftone did not vary considerably (Fig. 4). Following this example it can be seen how the image content influences the perception of quality.

The image content is filled with details and text. Higher contrast enables the FM raster to preserve the details of the image, and hue transitions hide the noise it creates. The distribution of raster dots with the AM halftone with fixed distance between the raster elements results in the loss of detail, which is why it is perceived as the lack of information, i.e. noise. As far as the rest of the samples are concerned, an increased number of observers that was unable to choose a

specific raster can be noticed, i.e. these observers did not see the difference between the different types of halftone on an image with this kind of content.

Table 5
Comparative analysis of different sizes of the test reproduction 3 for the parameter
the least noticeable grainy structure

	A (20 μm)%	B (40 μm)%	C (150lpi)%	D (175lpi)%	No difference%
3.1	63.4	6.9	13.9	9.9	5.9
3.2	63.4	4	13.9	10.8	7.9
3.3	41.5	5.9	23.8	23.8	5

χ^2=19.234, df=8, *p*-value: 0.01365708

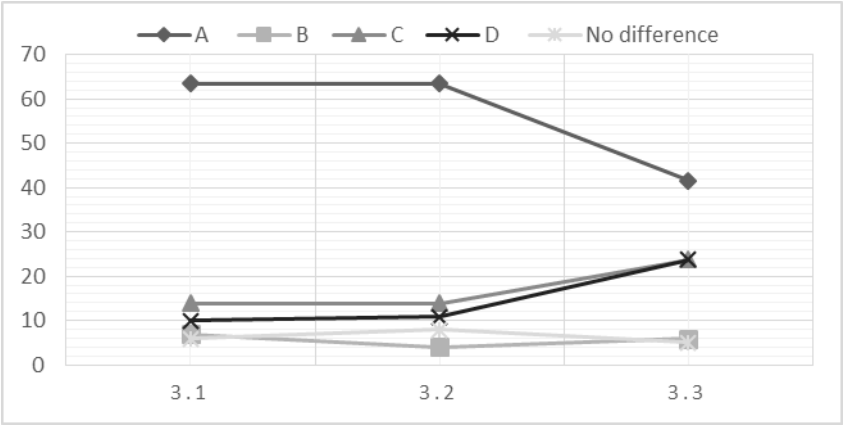


Figure 5
Observer's answers according to the parameter
the least noticeable grainy structure for different sizes of the test reproduction 3

In Table 5 percentages for the choice of FM decreased from 63.4% to 41.6% for the first two sizes with the largest reproduction, indicating increased perception of grainy structure on this size. Periodical type of halftone went up from 13.9% and, respectively, 10.9% to 23.8% on the largest reproduction. Test reproduction 3 has a combination of elements with plenty of detail and elements with smooth tonal gradations.

It can be seen, by analyzing the results, that with the increase of the image size a statistically significant difference, in the perception of the image, emerges. With smaller image sizes, FM halftone is dominant since it allows for a much better detail preservation (Fig. 5). Since the observed surface is smaller the noise on the image is not so obvious. With the largest test sample, a confrontation becomes inevitable between the elements with minute details and the surface with smooth tone transitions that start to show the signs of noise. This can be observed in the reduction of FM halftone percentage from 63.4% to 41.6%, and rising of

percentage for AM from 13.9% at 150lpi and 10.9% at 175lpi to 23.8% for each of the AM halftone, which is why one can say that AM raster takes over in dominance for the attribute of least noticeable grainy structure, at this image size.

Table 6
Comparative analysis of different sizes of the test reproduction 4 for the parameter *the least noticeable grainy structure*

	A (20 µm)%	B (40 µm)%	C (150lpi)%	D (175lpi)%	No difference%
4.1	85.1	5.9	3	4	2
4.2	84	3	5	5	3
4.3	88.1	5.9	2	1	3

χ^2=5.55, df=8, *p*-value: 0.69749434

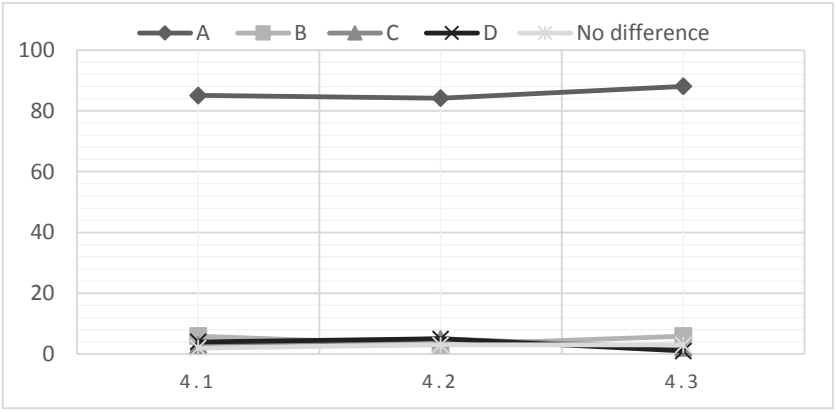


Figure 6
Observer's answers according to the parameter *the least noticeable grainy structure* for the different sizes of the test reproduction 4

Tables 6 and Figure 6 show that the car picture has no statistically significant differences in quality evaluation dependent on the size of the sample. Here the observers primarily choose FM 20 µm halftone. The cause can once again be found in the characteristics of both types of the screens.

Table 7
Comparative analysis of different sizes of the test reproduction 1 according to the parameter *the fidelity of reproduction*

	A (20 µm)%	B (40 µm)%	C (150lpi)%	D (175lpi)%	No difference%
1.1	50.5	3	9.9	35.6	1
1.2	39.5	3	19.8	34.7	3
1.3	6.9	7.9	34.7	49.5	1

χ^2=59.996, df=8, *p*-value: 0.00

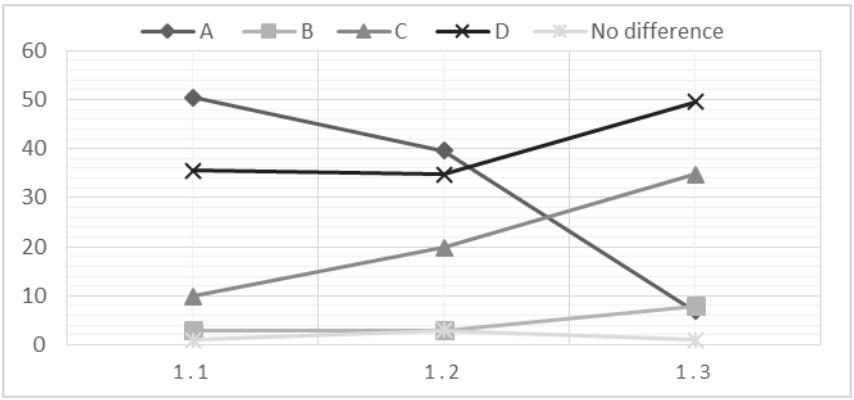


Figure 7
Observer's answers according to the parameter
the fidelity of reproduction for the different sizes of the test reproduction 1

On the smallest reproduction the expected distribution of the answers in favor of the 20 µm FM halftone can be seen, while already at the next size that percentage is reduced in the favor of 175lpi, which is, with the biggest reproduction, together with 150lpi AM, jointly they feature the highest fidelity of reproduction with over 80% (Table 7). Perceived quality of FM has dropped dramatically (Fig. 7), due to increasing in perceived graininess on images. Larger viewing angle, for this type of image content, enables uniform AM structures to be seen as reproduction with more fidelity than FM halftone.

The images that represent humans and faces are especially sensitive to the loss of detail. As one of the most common forms, each and every irregularity on a face or a human being is easily noticeable, hence the observers responded in such a way to the content of this image. The situation with the fidelity of the image is similar to the situation with the graininess. With the smallest size, the 20 µm FM raster is dominant while already at the medium size there is no statistically significant difference between the AM and FM. With the largest reproduction, the observers choose the AM halftone for the attribute *fidelity,* whereas the FM raster losses its fidelity because of the noise.

Table 8
Comparative analysis of different sizes of the test reproduction 2 according to
the parameter *the fidelity of reproduction*

	A (20 µm)%	B (40 µm)%	C (150lpi)%	D (175lpi)%	No difference%
2.1	52.6	7.9	6.9	16.8	15.8
2.2	53.5	4	10.9	25.7	5.9
2.3	65.3	4	13.9	9.9	6.9

χ^2=19.686, df=8, *p*-value: 0.011

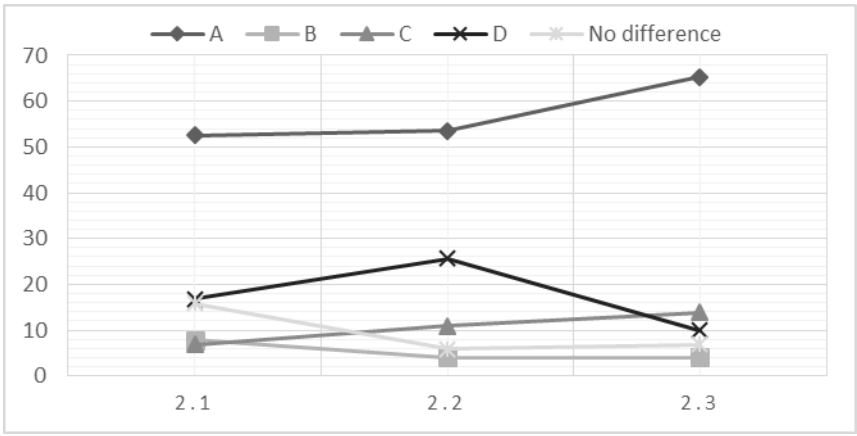


Figure 8
Observer's answers according to the parameter
the fidelity of reproduction for the different sizes of the test reproduction 2

On Figure 8, the distribution of the observer's answers for the reproduction 2 can be seen. Increasing the size of the image makes the FM halftone perceived as a high-fidelity image, with the choice frequency of 65.3%. It can also be seen (Table 8) that it held the greatest percentage without regard to the size of the test image, and with the increase of the size, the choice frequency also went up.

By analyzing the cross-tabulation data it can be noticed that the observers that choose the 175lpi AM halftoned image as the best for the middle size image, gave the advantage to the 20 μm FM halftone in the case of the largest image size. As far as 75.9% of those that gave the advantage to FM raster in the case of the middle size stayed true to their choice and chose it again in the case of the largest size. For other halftone types, it can be concluded that there is much indecisiveness on the part of the observers and that their answers differ according to the different sizes of the sample.

Table 9
Comparative analysis of different sizes of the test reproduction 3
according to the parameter *the fidelity of reproduction*

	A (20 μm)%	B (40 μm)%	C (150lpi)%	D (175lpi)%	No difference%
3.1	47.5	5.9	16.9	18.8	10.9
3.2	47.5	7.9	14.9	15.8	13.9
3.3	29.7	6.9	19.8	32.7	10.9

χ^2=13.924, df=8, *p*-value: 0.083

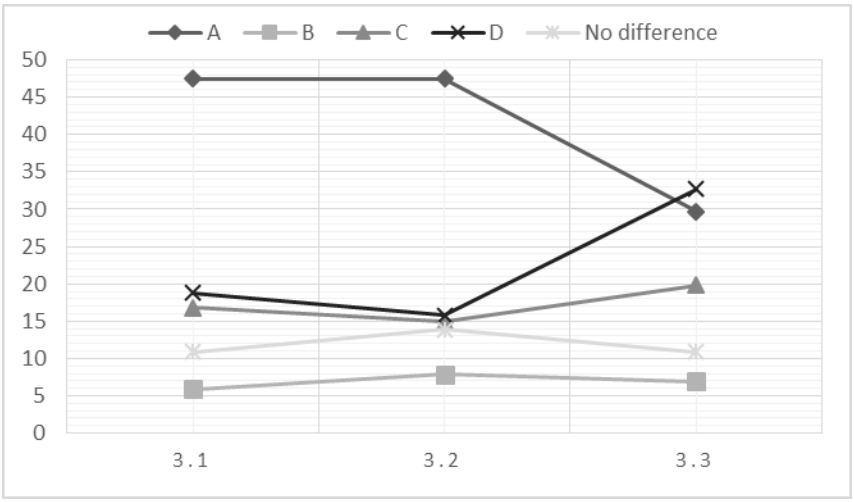


Figure 9
Observer's answers according to the parameter
the fidelity of reproduction for the different sizes of the test reproduction 3

Table 10
Comparative analysis of different sizes of the test reproduction 4
according to the parameter *the fidelity of reproduction*

	A (20µm)%	B (40µm)%	C (150lpi)%	D (175lpi)%	No difference%
4.1	69.3	4	12.2	9.9	4
4.2	62.4	5	11.9	12.9	7.8
4.3	60.4	7.8	14.9	12.9	4

χ^2=5.07, df=8, *p*-value: 0.750

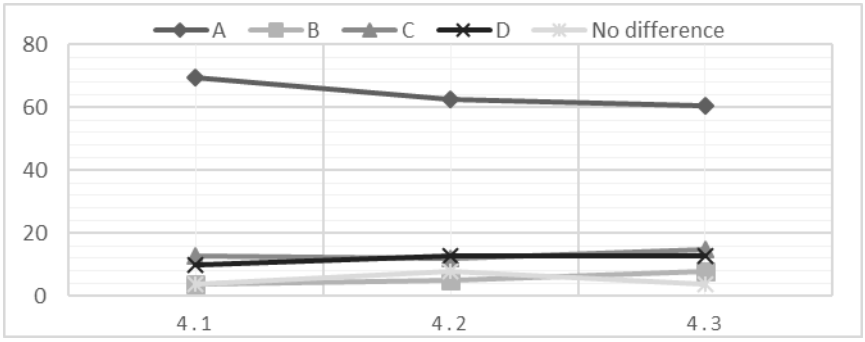


Figure 10
Observer's answers according to the parameter
the fidelity of reproduction for the different sizes of the test reproduction 4

Table 9 shows that test reproduction 3 has a statistically significant difference when examining high-fidelity attribute. On the smallest sizes, the FM halftone was marked as high-fidelity with 47.5%, while with the largest size the AM halftone was marked as the best with 32.7%, but the 20 μm FM was very close to AM with percentage of 29.7% (Fig. 9). By analyzing the cross-tabulation it can be seen that the largest number of the observers who marked the FM halftone as the best for the smallest and medium sizes of the test reproductions, chose the AM halftone as the best for the largest reproduction, which leads to the conclusion that there is a statistically significant difference in the choice of a specific halftone for the attribute of fidelity.

Figure 10 shows the very stable percentage of the observer`s choices, 20 μm FM halftone, without regard to the increase in the size of the reproduction. With the fourth image, there were no oscillations in the perception in the course of changing the size of the reproductions (Table 10). The picture consists of large surfaces with smooth tonal transition, which makes the observers choice interesting, since they choose FM with both parameters and all sizes as the high-fidelity reproduction. Halftone dots distributed in patterns are more pleasing to the eye of the observer when looking at uniform tone, which makes this result unexpected. The reason for this result can be also found in the definition of the parameters that the observers had to mark. Fidelity of the reproduction is a very broad term, which is synonymous with the universal image quality.

The research has shown that there is a statistically significant difference in the attributes of quality, *the least noticeable grainy structure,* and *the fidelity of reproduction*, in relation to the size of the image, which allows the H2 to be accepted.

Conclusion and Future Work

Considering the size of the halftone dot, viewing distance and conditions constant, and by varying only the size of the images, the conclusion can be reached that the perception of the halftone quality varies. Concerning *the least noticeable grainy structure*, an overall discussion of the results can be done. Combining the Tables 3-6 higher quality can be seen on smaller images, produced with 20 μm FM halftones. Tables 7-10 show that, regarding *fidelity,* FM halftone is less dominant but is still the best choice, when it comes to images with the smaller size. Increasing the size of the image will cause changes in the perception of reproduction quality, depending on the halftone and image content. On some content types, larger images show higher quality when they are rasterized by AM halftone. By analyzing the image content and the observer's answers, it can be seen that the biggest variation appears with the content that combines the elements of fine tone gradients and small details. The difference in the amount of information between the contone image and halftone image is reduced by increasing the size of the reproduction, which directly reflects on the increase of detail display with AM raster. The increase in the detail displayed by the AM

raster, correlates positively, with the perceived picture quality. With the highly detailed images having sharp contrast and dynamic range, the change in the size of the image influences the quality of the FM raster in a positive way, making it a recommended option in the case of images with such content. Based on the data analysis and other images of different content, it can be concluded that the change in the perceived quality of AM and FM halftone would happen, but only in the case of a more drastic size increase. Smaller images are more suitable for FM halftone reproduction. The threshold at which the AM rasterized image will be perceived as having better quality is dependent on the image content. Images with the combination of fine tone gradients and contrastive fine details have been shown to be the most sensitive to the size change. Further research is needed in order to establish a more precise connection between the amount of detail in the image, image size variation and the amount of data preserved in the image after the halftoning process. Including this data in the metric algorithms could provide new objective methods for determining the quality of printed reproduction. Comparison between AM and FM halftone is often done with subjective methods. Subjective research has a tendency to produce conflicting results, which is the reason the researchers often concluded that further research is needed to determine the reason for such oscillations. Conclusions of this research, stemming from the experiment, directly contribute to a better design of subjective assessments of print quality. It aids the researchers in obtaining the most consistent results possible and more clarified conclusions about different raster techniques.

Additional conclusions can be reached about the implementation of new raster techniques into the production system. Implementation of the new halftoning techniques carries different challenges, both technical and financial. The results of this paper can be of help in the process of weighing the possible benefits or negative results. By simple analysis of the graphic products that will be realized by the system, one can reach a more informed and profitable decision on which type of halftone to use. This decision can then be used in the course of implementation of a new halftoning technique in the desired graphic production system.

Use of both type of halftones on the press sheet provides the ability to exploit the positive characteristics of AM and FM. In order to fully exploit the potentials of raster techniques, the development of XM (Cross Modulated) algorithms should incorporate image content analysis. These are the attributes that can lead to an improvement of the distribution of AM and FM areas during the RIP process. This kind of approach would enable the choice threshold to contain more than just the gray level of halftone, but also additional parameters that would contribute to a better distribution of AM and FM halftones.

Acknowledgement

This research was supported by the Serbian Ministry of Science and Technological Development, Grant No.: 35027 "The development of software model for improvement of knowledge and production in graphic arts industry."

References

[1] Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P.: Image Quality Assessment: from Error Visibility to Structural Similarity, IEEE Trans. Image Process., Vol. 13, No. 4, Apr. 2004, p. 600

[2] Gooran S.: Hybrid Halftoning, a Useful Method for Flexography, J. Imaging Sci. Technol., Vol. 49, No. 1, 2005, p. 85

[3] Fung Y., Chan Y.: A Multiscale Error Diffusion Algorithm, in 17th European Signal Processing Conference (EUSIPCO 2009) Glasgow, 2009, p. 2258

[4] Liu H., Huang M., Liu Y., Wu B., Xu Y.: Color-Difference Evaluation for Digital and Printed Images, J. Imaging Sci. Technol., Vol. 57, No. 5, 2013, p. 1

[5] Dharavath H., Bensen T., Gaddam B.: Analysis of Print Attributes of Amplitude Modulated (AM) vs. Frequency Modulated (FM) Screening of Multicolor Offset Printing, J. Ind. Technol., Vol. 21, No. 3, 2005

[6] Karlović I., Tomić I., Jurič I., Pintier I.: Finding the Relation Between AM and FM Halftoning with S-CIE LAB Metrics, in 7th Symposium of Information and Graphic Arts Technology, Pardubice, 2014, p. 50

[7] Pinćjer I., Nedeljković U., Draganov S: Subjective Analysis of Image Quality : Experts and Naïve, in International Symposium on Graphic Engineering and Design GRID, Proceedings, Novi Sad, 2014, p. 449

[8] Ulichney R.: Review of Halftoning Techniques, Journal of Electronic Imaging, 1999, No. 1, p. 311

[9] Bayer B. E.: An Optimum Method for Two-level Rendition of Continuous-tone Pictures, IEEE International Conference on Communications, 1973, p. 11

[10] Velho, L., Gomes J.: Stochastic Screening Dithering with Adaptive Clustering, in SIGGRAPH '95 Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, Los Angeles, 1995, pp. 273

[11] Lau D. L., Arce G. R., Gallagher N. C.: Green-noise Digital Halftoning," Proc. IEEE, Vol. 86, No. 12, 1998, p. 2424

[12] Fung Y., Chan Y.: Tone-dependent Noise Model for High-quality Halftones J. Electron. Imaging, Vol. 22, No. 2, Apr. 2013, p. 023004

[13] He Z.: AM/FM Halftoning: Digital Halftoning Through Simultaneous Modulation of Dot Size and Dot Density, J. Electron. Imaging, Vol. 13, No. 2, Apr. 2004, p. 286

[14] N. Kašiković, D. Novaković, I. Karlović, and N. Milić, "Colourfastness of Multilayer Printed Textile Materials to Artificial Light Exposure," *Acta Polytech. Hungarica*, Vol. 12, No. 1, pp. 161-173, Feb. 2014

[15] Rovamo J. M., Kankaanpa M. I.: Modelling Spatial Contrast Sensitivity Functions for Chromatic and Luminance-modulated Gratings, Vision Res., Vol. 39, 1999, p. 2387

[16] Eerola T., Lensu L., Kälviäinen H., Kamarainen J.-K., Leisti T., Nyman G., Halonen R., Oittinen P.: Full Reference Printed Image Quality: Measurement Framework and Statistical Evaluation, J. Imaging Sci. Technol., Vol. 54, No. 1, 2010, p. 010201

[17] Wang Z., Simoncelli E. P., Bovik A. C.: Multiscale Structural Similarity for Image Quality Assessment, in The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, 2003, p. 1398

[18] Pedersen M., Bonnier N., Hardeberg J. Y., Albregtsen F.: Image Quality Metrics, Vol. 7867, Jan. 2011, p. 786702

[19] Streckel B., Steuernage B., Falkenhagen E., Jung E.: Objective Print Quality Measurements Using a Scanner and a Digital Camera, in DPP2003:IS&Ts International Conference on Digital Production Printing and Industrial Applications, Barcelona, 2003, p. 145

[20] Engeldrum P.: Extending Image Quality Models, IS TS PICS Conf., No. 5, 2002, p. 1

[21] Yu Q., Parker K. J.: Stochastic Screen Halftoning for Electronic Imaging Devices, J. Vis. Commun. Image Represent., Vol. 8, No. 4, Dec. 1997, p. 423

[22] Halonen R., Westman S., Oittinen P.: Naturalness and Interestingness of Test Images for Visual Quality Evaluation, in Proc. SPIE 7867, San Francisco, 2011

[23] Mahovic S., Mandić L., Agic D., Gojo M: A contribution to the AM and the FM Screening in the Graphic Reproduction Process, in DAAAM International Scientific Book 2005, Vienna, 2005, p. 395

[24] Itoua P., Beghdadi A., Lesegno P. V. D. E.: Objective Perceptual Evaluation of Halftoning using IQ Metrics, in 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010), Kuala Lumpur, 2010, p. 456

[25] Cui C.: Comparison of Two Psychophysical Methods for Image Color Quality Measurement: Paired Comparison and Rank Order, in Color and Imaging Conference, 8th Color and Imaging Conference Final Program and Proceedings, Scottsdale, 2000, p. 222

[26] Norberg O., Andersson M.: Perceived Image Quality of Printed Images and Their Relation to Paper Properties, in Seventeenth Color Imaging Conference: Color Science and Engineering Systems, Technologies, and Applications, Albuquerque, 2009, p. 210

[27] Engeldrum P.: A Theory of Image Quality: The Image Quality Circle, J. imaging Sci. Technol., Vol. 48, No. 5, 2004, p. 446

Process Network Solution of Extended CPM Problems with Alternatives

Nándor Vincze[1], Zsolt Ercsey[2], Tamás Kovács[3], József Tick[4], Zoltán Kovács[5]

[1] Department of Applied Informatics, Faculty of Education, University of Szeged, Boldogasszony u. 6, 6725 Szeged, Hungary, vincze@jgypk.u-szeged.hu

[2] Department of System and Software Technology, Faculty of Engineering and Information Technology, University of Pécs, Boszorkány u. 2, 7624 Pécs, Hungary, ercsey@mik.pte.hu

[3] Department of Computer Algorithms and Artificial Intelligence, Institute of Informatics, University of Szeged, Árpád tér 2, 6720 Szeged, Hungary, tamas.kovacs@optin.hu

[4] Department of Applied Informatics, John von Neumann Faculty of Informatics, Óbuda University, Bécsi út 96/b, 1034 Budapest, Hungary, tick@uni-obuda.hu

[5] Department of Computational Optimization, Institute of Informatics, University of Szeged, Árpád tér 2, 6720 Szeged, Hungary, kovacsz@inf.u-szeged.hu

Abstract: In this paper a novel method to extend the problem range of CPM problems is given. First, the CPM problem is transformed into a process network problem. It is shown how the directed bipartite process network has to be generated, and the corresponding mathematical programming model should be formulated. Time optimal, cost optimal, time optimal with additional cost constraints and cost optimal with additional time constraints mathematical programming models are given. Moreover, it is illustrated how alternative cases may appear in the structure. An example illustrates the efficiency of the present work.

Keywords: CPM; process network; alternatives

1 Introduction

The critical path method (CPM) is an algorithmic approach of scheduling a set of activities. CPM is widely used for projects in the field of constructions to software

development. Modeling techniques originate back in the 1950s. The main criteria, in order to use the CPM technique, are the following. First, duration times of the activities have to be known together with the dependencies between the activities. Based on this information the activity network is developed. With the help of the list of activities together with their duration and dependencies on each other as well as on the logical end points, CPM calculates the longest path of the planned activities together with the earliest and latest times that each activity can start or finish without lengthening the project. In this regard, the critical path is the sequence of the activities which add up to the longest overall duration. Please note that there are activity-on-node and activity-on-arc approaches of the CPM from which representations the latter is considered for the present work. For further information please consider the problem definition of Chanas and Zielinski, 2001.

Recent advances consider CPM networks with ant colony optimization methods (Shankar et al. 2011, Li Hui et al. 2013); CPM problems where task durations are uncertain (Li et al. 2015), fuzzy linear programming model (Madhuri et al. 2013). For a special problem CPM with time–cost trade-offs, i.e. process of crashing, was investigated by Sunita et al. 2013, while activity list-based nested partitions algorithm with local adjustments were used by Xiao et al. 2014.

Please note that CPM techniques order the resources to the activities, but this order is not represented in the graph. When another resource is ordered to the same activity then the parameters of the problem have to be reset, in most cases a new graph has to be depicted, and the problem has to be solved again; i.e. a large number of separate problems has to be uniquely considered. Moreover, CPM graph techniques do not handle at all cases where a given subtask can be solved in many different ways.

Friedler et al. 1992a, 1992b, introduced a process network methodology for chemical engineering problems. Based on rigorous mathematical foundations the approach relies both on graph theory as well as combinatorial techniques focusing first on the structure generation of the problem considered. Besides the directed bipartite process network an underlying axiom system is used to derive theorems to generate the potentially feasible structures as well as the so-called maximal structure which includes all feasible solution structures. When the algorithmically and mathematically proven structure generation ends, a mathematical programming model is generated and solved with similar mathematical rigor. One of the main advantages of the developed methodology is that alternative solutions can also be interpreted for the problems considered.

Later on this methodology was used in other fields, for example workflow modelling (Tick 2007, Tick et al. 2013); separation network synthesis problems with multiple feed streams and sharp separators (Kovács et al. 1999 and 2000); generating and analyzing structural alternatives for supply scenarios (Barany et al. 2010; Klemeš et al. 2010, Kalauz et al. 2012); determining the thermodynamically dominant pathways in a metabolic network (Yun et al. 2013), or identifying

feasible pathways of the reaction catalysed by a catalyst with multi-active sites (Fan et al. 2012). Process networks were successfully adopted for solving the routing and scheduling of evacuees, facing a life-threatening situation as well (Garcia-Ojeda et al. 2012).

Please note that there has been no connection between CPM and process network methodology in the literature until today.

1.1 Mapping of CPM to a Process Network

In order to solve CPM problems with the help of process network methodology, the two terminologies have to be mapped. First, the basic elements are considered as described in Table 1. Both the CPM and the process network methodology may have attributes corresponding to their various objects, depending on the application field. For example, in the CPM an activity has a given duration and may have various resources with various costs; while on the other hand in a process network an operating unit may also have a given operation time and various costs.

Table 1
Basic terminology of the CPM and of the process networks

CPM	Process network
Event (node). Activity (arc). Logical connection between the activities (dependencies between the activities). CPM graph: scheduling problem.	Material (node type 1): raw material, intermediate and product. Operating unit (node type 2). Material flow (arc). Process network: network of the operating units producing the products from the raw materials.

It is worth mentioning that in the CPM networks the work is performed by the resources, while in the process networks work is performed by the operating units. Therefore, these are mapped to each other.

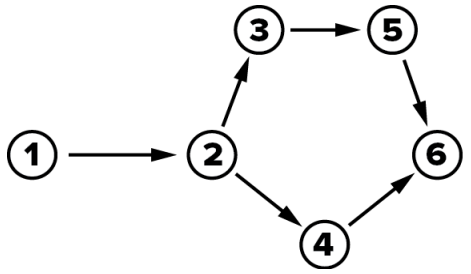


Figure 1a
CPM graph example

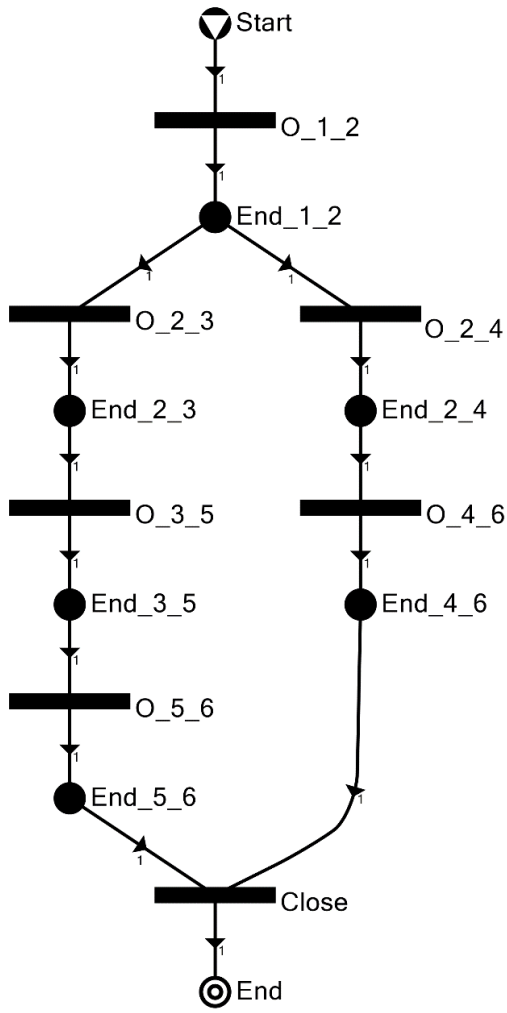


Figure 1b

Process network of the CPM graph example A given CPM graph can be mapped into a process network as follows. There is a given start of the CPM problem, see Figure 1a. In the process network it is represented by a raw material, see "Start" in Figure 1b. The activity between event 1 and event 2 of the CPM graph is represented in the process network as operating unit O_1_2. The input material of the operating unit O_1_2 in this case is the "Start." Activity between event 2 and event 3 of the CPM starts when event 2 is accomplished. In the process network it is mapped into the material End_1_2, which is the output material of O_1_2, which means that the operation of O_1_2 is finished. Generally speaking, in the process network such materials will be inputs to an operating unit, which have to be accomplished in the CPM graph before the activity can start. In the process

network of the current example operating unit Close will have two input materials, namely End_4_6 and End_5_6, since in the CPM graph both event 4 and event 5 have to be accomplished before the activity between 6 and 7 can start. Each operating unit of the process network has one output, which represents the operational finish of the operating unit. Since the end point of the CPM graph is event 6 with two different preceding activities, in the process network an addition technical operating unit, called Close, has to be inserted. As a result, the CPM graph of Figure 1a can be mapped into the process network Figure 1b.

When the structural mapping is done according to the above given details, Table 2 illustrates the logical connections between the CPM and process networks.

Table 2
Logical connection between the CPM and process networks

CPM	Process network
Activity.	Operating unit.
Event.	Material (raw material, intermediate and product)
Logical connection between the activities (dependencies between the activities).	Material flow (arcs).
CPM graph: scheduling problem.	Process network: network of the operating units producing the products from the raw materials.

The mapping is done accordingly, then the basic terminology of CPM and process networks can be combined for simplicity.

1.2 Illustrative Example of the Mapping

Let us consider the example illustration published by Chanas and Zielinsky (2001). The CPM graph of the example is given in Figure 2a; while Figure 2b illustrates the process network representation of the example after the mapping. Please note that in Figure 2a, the activity between event 4 and event 7 starts only when both activities between event 2 and event 4 and between event 3 and event 4 are finished. Similarly, in Figure 2b operating unit O_4_7 starts its operation only when both operating units O_2_4 and O_3_4 finished their operations, i.e. materials End_2_4 and End_3_4 are available respectively, serving as input materials of O_4_7. Please also note that since the operating units under consideration are different, their end points are also different.

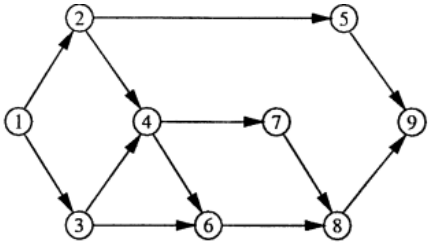


Figure 2a
CPM graph of the illustrative example

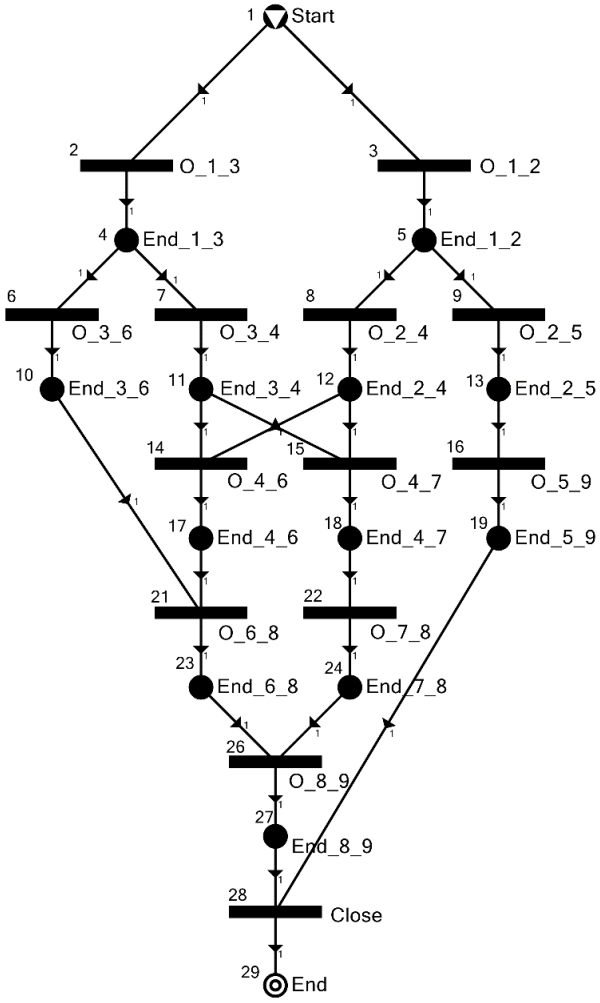


Figure 2b
Process network of the illustrative example

2 Alternatives

It is important to note that real case examples raise the question of alternatives. In case a given problem can be solved by performing more than one activity or more than one series of activities, then it is called a problem of alternatives. This situation is not handled by CPM, moreover, crucial decisions in this regards have to be made prior to the depiction of the CPM graph. Nevertheless, these decisions may fundamentally influence the overall duration of the final result of the CPM solution. Obviously, it would be of high importance not to exclude any possibilities at the beginning, but to have these decisions as a result of a solution process as well. The transformation presented in this paper gives the possibility of making these decisions later on. Alternatives can be added into the process network and thus CPM problems extended with alternatives can be considered within this framework. In the process network representation of the problem these are considered as alternative arcs. This can be easily illustrated in the process network by adding a parallel operating unit for the alternative group, see Figure 3.

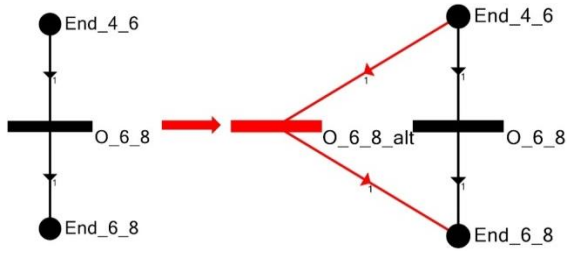


Figure 3
Illustration of adding an alternative

Based on the mapping of the basic elements and logical connections of the CPM and process network representations described earlier in this paper, it is obvious that the problem of alternative arcs and alternative paths can be described as alternative operating units of the process network. As a result, project processes can be illustrated in more details in the process network representation than in CPM graphs. A given activity is performed by one operating unit or another, or one phase of the work is performed by one operating unit and another phase by another operating unit. This can be imagined for example when during a production process a decision maker controls the alternative operating units (arcs), namely which operating unit should operate during the work process.

In the process network representation, this means information additional to the structural representation. Until this point structural construction was performed, in other words from the raw materials and set of operating units the production process was generated to produce the desired products with the help of additional intermediate materials. Here, a decision maker may set up priorities and may decide between alternatives. Normally, this information is represented in the cost function of the operating units.

All in all, based on the mapping described previously, the CPM graph was transformed into a process network and then alternative solutions were also added into the process network representation, see alternative operating units O_6_8_alt and O_8_9_alt in Figure 4. Please note that the input and output materials of the added alternative operating units are identical to that of the original operating units.

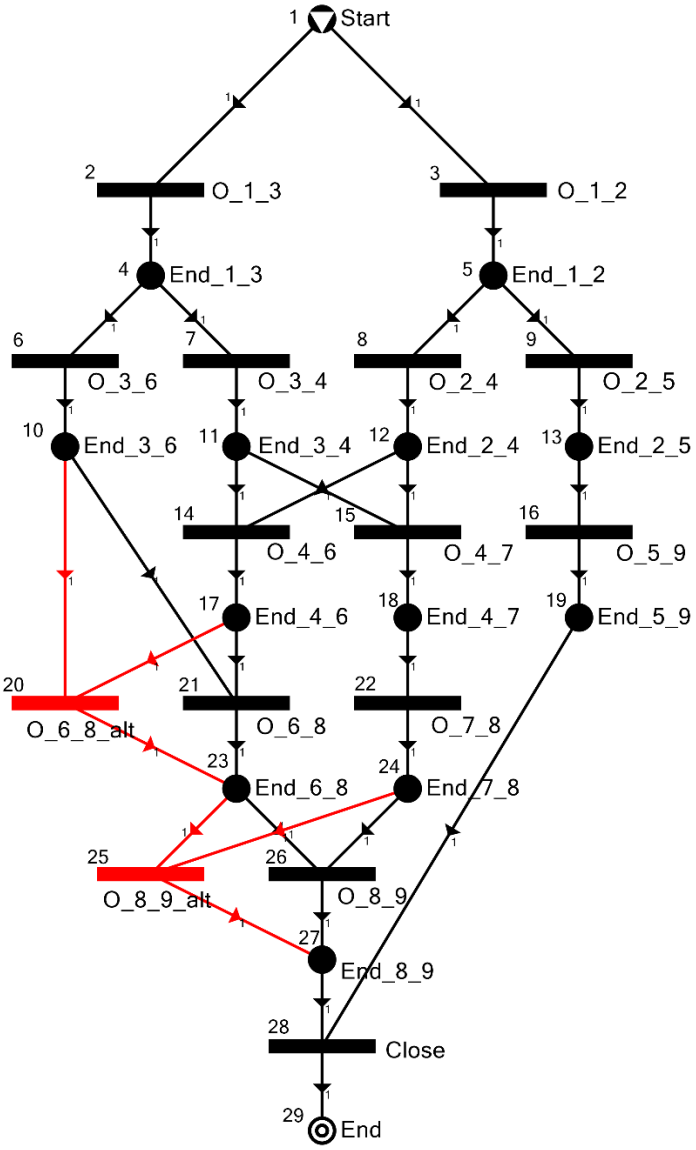


Figure 4
Illustrative example with alternatives

The axioms of process networks (see Friedler et al. 1992a) determine a solution structure within the process network. According to the terminology of the CPM, the axioms of process networks have to be extended with the following: each event in any solution structure, which is represented by a material in the process network, has one and only one input arc, except the Start event, which has zero indegree. In other words, it is also important to lock out from the solution point of view the parallel alternatives. Therefore, it has to be stated that from every operating unit there exists one and only one path to the final product (which is the end of the project in case of the CPM model).

In this regards, the solution structures of the process network now correspond to the CPM graph. Therefore, adding the alternatives within the process network, multiple CPM graphs are described for the original problem. As a result, the optimal solution with a given set of constraints of the original problem is generated from the mathematical programming model of the process network with alternatives. The proposed solution method considering the alternatives is also given in Figure 5.

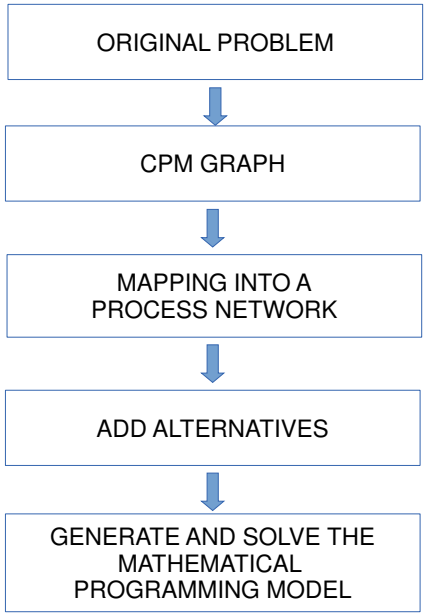


Figure 5
Proposed solution method

3 Mathematical Programming Model

Let *A*, *E*, *V* be finite sets, where *A* denotes the set of Activities, *E* denotes the set of Events and *V* denotes the set of vertices. Let *G* be the bipartite process network as follows.

$G(A,E,V);\ A \cap E = \varnothing, V \subseteq \left(A \times E\right) \cup \left(E \times A\right)$ *bipartite graph*

$A = \left\{ i \in N \right\}$ *activities*

$E = \left\{ j \in N \right\}$ *events*

As an illustration, the above formula means in Figure 4 the followings:

A={2,3,6,7,8,9,14,15,16,20,21,22,25,26,28}

$E = \{1,4,5,10,11,12,13,17,18,19,23,24,27,29\}$

$V = \{(1,2),(1,3),(2,4),(3,5),\ldots\}$

Let us denote x_i the *i*-th activity in the CPM graph and operating unit in the process network, where

$$x_i = \begin{cases} 0 & \textit{the i-th activity is not performed} \\ 1 & \textit{the i-th activity is performed} \end{cases}$$

t_i = the time from start-up to the *i*-th event occurs.

T_i = the duration of the *i*-th activity.

T = planned upper time limit for the total project.

C_i = cost of the *i*-th activity.

C = planned upper budget for the total project.

$x_{Close} = 1$	$Close \in A$	(1)
$\sum_{\{i:\, i \in A \text{ and } (i,j) \in V\}} x_i = 1$	$\forall j \in E \text{ and } j \neq Start$	(2)

Line (1) refers to the fact that the project has to be finished; line (2) refers to that only one alternative can be considered. Let $\left\{ b_1, b_2, ..., b_n \right\}$ be the solution of the above equations and let $S = \left\{ i : i \in A \text{ and } b_i = 1 \right\}$ be a subset of *A*. The *S* set of nodes and the corresponding edges designates a part of G(A, E, V) process graph, which exactly specify a CPM graph.

As an illustration, all CPM graphs of the illustrative example depicted in Figure 4 are as follows:

CPM graph 1: $b_i = \begin{cases} 0 & if\ i = 20, and\ i = 25 \\ 1 & otherwise \end{cases}$ where $i \in A$;

CPM graph 2: $b_i = \begin{cases} 0 & if\ i = 21, and\ i = 26 \\ 1 & otherwise \end{cases}$ where $i \in A$;

CPM graph 3: $b_i = \begin{cases} 0 & if\ i = 20, and\ i = 26 \\ 1 & otherwise \end{cases}$ where $i \in A$;

CPM graph 4: $b_i = \begin{cases} 0 & if\ i = 21, and\ i = 25 \\ 1 & otherwise \end{cases}$ where $i \in A$.

3.1 Mathematical Programming Model of the Time Optimal Project Plan

The mathematical programming model for the usual CPM problem, extended with alternatives is given below.

$$x_{Close} = 1 \qquad Close \in A \tag{3}$$

$$\sum_{\{\, i:\, i \in A \text{ and } (i,j) \in V \}} x_i = 1 \qquad \forall j \in E \text{ and } j \neq Start \tag{4}$$

$$t_{Start} = 0 \qquad Start \in E \tag{5}$$

$$t_k + \sum_{\{\, i: i \in A \text{ and } (i,j) \in V \}} x_i T_i \leq t_j \qquad \forall k, j \in E \setminus Start \text{ where } \exists i\ (k,i) \text{ and } (i,j) \in V \tag{6}$$

$$\overline{t_{End} \rightarrow \min} \tag{7}$$

Please note that lines (3) and (4) are equal to lines (1) and (2) above; while line (4) refers to that only one alternative can be considered, line (5) refers to the fact that the project started at time zero, line (6) indicates that the activities have been completed at the earliest time. Line (7) refers to the aim, which is to minimize the overall duration of the project.

Please also note that the solution of this mathematical programming model equals to the solution in which the shortest duration time is chosen among the alternatives.

3.2 Mathematical Programming Model of the Cost Optimal Project Plan

The mathematical programming model for the resource allocation problem of the given CPM graph, extended with alternatives is given below:

$$x_{Close} = 1 \qquad Close \in E \tag{8}$$

$$\sum_{\{i:\, i \in A \text{ and } (i,j) \in V\}} x_i = 1 \qquad \forall j \in E \text{ and } j \neq Start \tag{9}$$

$$\overline{\sum_{\{i:\, i \in E\}} x_i C_i \rightarrow \min} \tag{10}$$

Please note that lines (8) and (9) are equal to lines (1) and (2) above. Line (10) refers to the aim, which is to minimize the overall cost of the project.

3.3 Mathematical Programming Model of the Time Optimal Project Plan with Additional Cost Constraint

In real case examples, not only the time constraints are important for the project planning but also the different costs of the alternatives are also of importance. Thus the previously described mathematical programming model can be reformulated as follows:

$$x_{Close} = 1 \qquad Close \in E \tag{11}$$

$$\sum_{\{i:\, i \in A \text{ and } (i,j) \in V\}} x_i = 1 \qquad \forall j \in E \text{ and } j \neq Start \tag{12}$$

$$t_{Start} = 0 \qquad Start \in E \tag{13}$$

$$t_j \geq x_i T_i + t_k \qquad \forall j \in E \setminus Start \text{ and } \forall i : (i,j) \in V \text{ and } \forall k : (k,i) \in V \tag{14}$$

$$\sum_{\{i:\, i \in A\}} x_i C_i \leq C \tag{15}$$

$$\overline{t_{End} \rightarrow \min} \tag{16}$$

Please note that lines (11) and (12) are equal to lines (1) and (2) above; while line (13) refers to the fact that the project started at time zero, line (14) indicates that the activities have been completed at the earliest time, while line (15) refers to the fact that the total project cost should not exceed the given upper budget limit. Line (16) refers to the aim, which is to minimize the overall duration of the project.

3.4 Mathematical Programming Model of the Cost Optimal Project Plan with Time Constraint

Finally, the mathematical programming model of the cost optimal project plan with time constraint can be given as follows:

$$x_{Close} = 1 \qquad Close \in A \tag{17}$$

$$\sum_{\{i:\, i \in A \text{ and } (i,j) \in V\}} x_i = 1 \qquad \forall j \in E \text{ and } j \neq Start \tag{18}$$

$$t_{Start} = 0 \qquad Start \in E \tag{19}$$

$$x_i T_i + t_k \leq t_j \qquad \forall j \in E \setminus Start \text{ and } \forall i : (i,j) \in V \text{ and } \forall k : (k,i) \in V \tag{20}$$

$$t_{End} \leq T \qquad End \in E \tag{21}$$

$$\sum_{\{i:\, i \in A\}} x_i C_i \rightarrow \min \tag{22}$$

Please note that lines (17) and (18) are equal to lines (1) and (2) above; while line (19) refers to the fact that the project started at time zero, line (20) indicates that the activities have been completed at the earliest time, while line (21) refers to the fact that the total project duration should not exceed the given upper time limit. Line (22) refers to the aim, which is to minimize the overall cost of the project.

4 Illustrative Example

The illustrative example from Chanas S. and P. Zielinski (2001) is revisited below, for details please consider the process network given in Figure 4. Please note that the given fuzzy times of the example were first transformed to triangular fuzzy times and then they were defuzzificated to the crisp times below. Thus, the triangular fuzzy time and the defuzzificated crisp times are indicated in Table 3.

Table 3
Duration of the *i*-th activity

Triangular fuzzy times (CPM)	**Defuzzificated crisp times (process network)**
T_{12}=(1,25;1;1)	T_3=1,25
T_{13}=(2,5;1;1)	T_2=2,5
T_{24}=(0;0;0)	T_8=0
T_{25}=(2,5;1;1)	T_9=2,5
T_{34}=(0;0;0)	T_7=0
T_{36}=(6,5;1;1)	T_6=6,5

T_{46}=(5;1;1)	T_{14}=5
T_{47}=(9;1;1)	T_{15}=9
T_{59}=(8,5;2;2)	T_{16}=8,5
T_{68}=(4;2;2)	T_{21}=4
T_{78}=(3,5;1;1)	T_{22}=3,5
T_{89}=(7,5;2;2)	T_{26}=7,5

The mathematical programming model of the time optimal project plan with additional cost constraint given above in line (11) – (16) was formulated. Let us consider the following costs of the activities given as follows:

$$\left\{\begin{array}{l} C_2 = 5, C_3 = 5, C_6 = 5, C_7 = 5, C_8 = 5, \\ C_9 = 5, C_{14} = 5, C_{15} = 5, C_{16} = 5, \\ C_{21} = 10, C_{22} = 5, C_{26} = 10, C_{28} = 5 \end{array}\right\};$$

with the duration of the alternative activities: T_{20}=8; T_{25}=9,5; and the costs of the alternative activities: C_{20}=5; C_{25}=5. Let us consider the planned upper budget for the total project C = 80. Thus, the solution of the mathematical programming model results in the overall project duration is 22,5; and in this optimal solution, nodes 21 and 26 are selected, and the alternatives are excluded.

Should the planned upper budget for the total project be C = 70; then the solution of the mathematical programming model results in the overall project duration 24,5; while in this optimal solution, alternative node 20 is selected instead of the original 21, and the alternative node 25 is excluded.

Conclusions

The critical path method (CPM) gives the longest path of the planned activities together with its overall duration. In each case only time is considered and only one solution is found at a time. Resources together with their costs do not appear in CPM graphs.

In the present paper a novel method to extend the problem range of CPM problems is given. First, it was illustrated how a CPM problem should be transformed into a process network problem. This mapping has no literature antecedents since this situation is not handled in the CPM graph, moreover there is no CPM and process network connection yet. This transformation serves as the key to commonly handle resources together with their costs already within the considered process network. Since real case examples raise the question of alternatives, namely when more than one activity or more than one series of activities are considered for a subtask, this can also be represented in the process network.

CPM methods order the resources to the activities only without any influence to the CPM graph. When another resource is considered to be used to the same activity then in most cases all the parameters have to be reset and a new graph has

to be depicted, and thus the problem has to be solved again. This means a large number of problems to be solved separately, and the parameters considered for each separate problem are independent from the other separate problems' parameters. However, these parameters are dependent on each other in real case problems. Moreover, CPM graph techniques do not handle at all such cases where a given problem can be solved by performing more than one activities, i.e. alternatives. In other words crucial decisions regarding alternatives have to be made before the CPM graph is depicted. In the present work as a novel solution alternatives can be added and are handled within the given model. Moreover, when alternatives are added to the process network all resources together with their parameters depending on each other are considered and handled within the model, as in the case of real case problems.

After describing the extended problem with alternatives, four different mathematical programming models are given in the present work. These mathematical programming models cover a wider range of problems than that of the CPM methodology. These mathematical programming models can be generated and solved algorithmically. The solution of the mathematical programming model is the optimal solution of the original CPM problem, i.e. the exact project definition, solution. Another novelty of the present work is that all solutions of the original CPM problem can be generated and ranked in order according to the cost function.

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] Barany M., Bertok B., Kovacs Z., Friedler F., Fan L. T., Optimization Software for Solving Vehicle Assignment Problems to Minimize Costs and Environmental Impacts of Transportation, Chemical Engineering Transactions, (2010) 21, 499-504

[2] Chanas S. and P. Zielinski, Critical Path Analysis in the Network with Fuzzy Activity Times, Fuzzy Sets and Systems, (2001) 122, 195-204

[3] Fan, L. T., Y.-C. Lin, S. Shafie, B. Bertok, and F. Friedler, Exhaustive Identification of Feasible Pathways of the Reaction Catalyzed by a Catalyst with Multiactive Sites via a Highly Effective Graph-Theoretic Algorithm: Application to Ethylene Hydrogenation, Industrial & Engineering Chemistry Research, (2012) 51(6), 2548-2552

[4] Friedler, F., K. Tarjan, Y. W. Huang, and L. T. Fan, Graph-Theoretic Approach to Process Synthesis: Axioms and Theorems, Chem. Engng Sci., (1992a)°47, 1973-1988

[5] Friedler, F., K. Tarjan, Y. W. Huang, and L. T. Fan, Combinatorial Algorithms for Process Synthesis, Computers Chem. Engng, (1992b) 16, S313-320

[6] Garcia-Ojeda, J. C., B. Bertok, F. Friedler, Planning Evacuation Routes with the P-Graph Framework, Chemical Engineering Transactions, (2012) 29, 1531-1536

[7] Kalauz K., Sule Z., Bertok B., Friedler F. and Fan L. T., Extending Process-Network Synthesis Algorithms with Time Bounds for Supply Network Design, Chemical Engineering Transactions, (2012) 29, 259-264

[8] Klemeš J., Friedler F., Bulatov I., Varbanov P., Sustainibility in the Process Industry: Integration and Optimization (Green Manufacturing & Systems Engineering), McGraw-Hill Professional, New York, USA (2010)

[9] Kovacs Z, Z Ercsey, F Friedler and L. T. Fan, Exact Super-Structure for the Synthesis of Separation-Networks with Multiple Feed-Streams and Sharp Separators; (1999) Computers and Chemical Engineering 23:(Supplement 1) pp. S1007-S1010

[10] Kovacs Z, Ercsey Z, Friedler F, Fan L T Separation-Network Synthesis: Global Optimum through Rigorous Super-Structure (2000) Computers and Chemical Engineering 24:(8) pp. 1881-1900

[11] Yun, C., T. Y. Kim, T. Zhang, Y. Kim, S. Y. Lee, S. Park, F. Friedler, and B. Bertok, Determination of the Thermodynamically Dominant Metabolic Pathways, Industrial & Engineering Chemistry Research (2013) 52(1), 222-229

[12] József Tick, P-Graph-based Workflow Modelling, Acta Polytechnica, Vol. 4, No. 1 (2007) 75-88

[13] József Tick, Csanád Imreh, Zoltán Kovács, Business Process Modeling and the Robust PNS Problem, Acta Polytechnica, DOI: 10.12700/APH.10.06.2013.6.11, Vol. 10, No. 6 (2013) 193-204

[14] Zhenhong Li, Yankui Liu, Guoqing Yang: A New Probability Model for Insuring Critical Path Problem with Heuristic Algorithm. Neurocomputing 148: 129-135 (2015)

[15] Li Hui, Zhang Jingxiao, Ren Lieyan, Shi Zhen, Scheduling Optimization of Construction Engineering based on Ant Colony Optimized Hybrid Genetic Algorithm. Journal of Networks, Vol. 8, No. 6, June 2013, pp. 1411-1416

[16] Madhuri, U., Saradhi, P. ve Shankar, R., (2014) "Fuzzy Linear Programming Model for Critical Path Analysis", Int. J. Contemp. Math. Sciences, Vol. 8, No. 2, 2013, pp. 93-116

[17] Shankar, N. Ravi, P. Phani Bushan Rao , S. Siresha and K. Usha Madhuri, Critical Path Method in a Project Network using Ant Colony Optimization.

International Journal of Computational Intelligence Research ISSN 0973-1873, Vol. 7, No. 1 (2011), pp. 7-16

[18] Sunita, K and Snigdha, B., CPM Analysis of Rolai-Rinjlai Road Construction, Research Journal of Mathematical and Statistical Sciences. Vol. 1, No. 2, 2013, pp. 7-15

[19] Xiao, L., Tian, J., & Liu, Z. (2014 June) An Activity-List based Nested Partitions algorithm for Resource-Constrained Project Scheduling. In Intelligent Control and Automation (WCICA), 2014 11th World Congress on (pp. 3450-3454) IEEE

The Influence of Demographics, Job Characteristics and Characteristics of Organizations on Employee Commitment

Valentin Kónya[1], Dejan Matić[2], Jasmina Pavlović[2]

[1] Faculty of Economics, University of Novi Sad, Segedinski put 9-11, 24000 Subotica, Serbia

[2] Department of Industrial Engineering and Management, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia

valentink@uns.ac.rs, dejan.matic@uns.ac.rs, jasmina.pavlovic@uns.ac.rs

Abstract: How individuals behave in organizations became an emerging topic among both the scientific and business world during the past decade. Commitment of employees to their work and the organization is considered an important indicator of successful organizational behavior. Previous research on the demographic and individual characteristics of employees, as well as, job and characteristics of organizations has shown that they have significant influence on organizational commitment. This paper intends to reveal some crucial information on how these characteristics influence organizational commitment in Central European organizations. Our analysis resulted in several important findings, revealing some interesting differences and when compared to other studies conducted in mostly western developed environments:

1) Gender does not have any influence on organizational commitment

2) Characteristics of organizations and most demographic characteristics have little effect on organizational commitment

3) Job characteristics have strong impact on organizational commitment in Central European Organizations

The research, in this paper, is part of a broader study, investigating the mutual influence of several leadership and organizational behavior related variables. Additional, partial results of this study, published in papers and conference proceedings from 2014, with the overall results scheduled for publishing in 2015.

Keywords: organizational commitment; demographic characteristics of employees; job characteristics; characteristics of organizations

1 Introduction

Problems that have emerged in contemporary business environments are related to increased market demands, increased unemployment, quality and service improvements, demand for new knowledge and skills, as well as the need for innovation and creativity of employees. All these problems contributed to increased importance concerning how individuals behave in organizations. The essential questions related to employees are how to attract high quality staff, how to motivate them to achieve top of the notch results and how to keep them in the organization?

As a factor of employee motivation, the commitment level of employees to their work and their organization is considered one of the most important indicators for a successful organizational behavior. Employees with higher levels of commitment are devoted to their professions and the organization, expect high demands from themselves, achieve superior results and demonstrate superior work performance.

Previous research on the demographic and individual characteristics of employees has shown that they are related to organizational commitment. It has been determined that a positive linkage exists between the age of individuals, years they have spent in an organization and the level of their commitment. Further, subjective acceptance of an organization, in the form of a psychological contract between employees and the organization, is of a great importance for building and gaining organizational commitment among employees.

In complex and continuous transitional conditions, it is of great importance for the processes of work and professional education, selection and employee development as well as organizational management strategies, to determine the factors influencing commitment to work and organization.

As two post-communist states, Serbia, with its Central European province of Vojvodina and Hungary were chosen for our research due to their transitional and post-transitional economies, as well as, both the public and private organizations on which our research was focused.

2 Organizational Commitment

Commitment in organizations is, most of the time, studied as an individual level variable within the framework of Organizational Behavior. However, it can also be viewed as a group or team level variable as the commitment of groups and teams. Another differentiation of commitment is related to the context of the object to which employees are committed, in the form of their feelings and beliefs.

The most often cited object of commitment, are whole organizations, i.e. organizational commitment [1] and this area of commitment research developed first [2]. Commitment, is often, related to feelings and beliefs about certain units inside and outside of organizations, such as work, job, team, group, association, union, profession, etc. [3]. Some authors make common errors in differentiating organizational and job commitment, often viewing the two as a single category. These two indeed have noticeable common points and overlaps, but surely are separate categories.

Organizational commitment is a work related attitude [4, 5]. Because attitudes influence our behavior toward objects, situations, persons or groups, the simplest way to define organizational commitment is to say it is an attitude that reflects the strength of relations between an organization and its employees [6], or the extent to which an employee is loyal to his or her organization [7]. Kanter [8] was one of the first do define commitment as the willingness of a social actor to give his/her energy and loyalty to a social system. In terms of organizational commitment, the term actor refers to employees and the term system refers to an organization. Porter et al. [4, p. 604] defines organizational commitment as "a strong belief in and acceptance of the organization's goals, a willingness to exert considerable effort on behalf of the organization, and a definite desire to maintain organizational membership". Similarly, Bateman and Strasser [9, p. 95] discuss that commitment is an organizational variable "involving an employee's loyalty to the organization, willingness to exert effort on behalf of the organization, degree of goal and value congruency with the organization, and desire to maintain membership". Rusbult and Farrel [10, p. 430] view commitment as "the likelihood that an individual will stick with a job and feel psychologically attached to it, whether it is satisfying or not".

Two types of organizational commitment are studied in this paper: affective commitment and continuance commitment. The first is the commitment to the values of the organization, also referred to as affective commitment. Commitment to the values of the organization is an emotional commitment to the organization, identification with the organization and its values and goals, as well as engagement in the organization. Employees that are committed to organizational values of their organizations do not think about leaving: they stay in their organizations because they want to, with their own free will [3, 11-13]. This type of commitment is characterized with strong belief in the organization, accepting its goals and values, readiness to exert extra effort and hard work in favor of the organization and finally, definitive desire to keep organizational membership [4]. The predictors of commitment to the values of the organization are the personal characteristics and working experiences of employees, while the confirmed benefits are reduced fluctuations and absenteeism, higher level of attendance to work and work performance, organizational citizenship behavior (OCB) and positive impact on employee health [13, 14].

The second type is commitment to stay in the organization or continuance commitment, related to the perceived costs of leaving the organization. Employees expressing high level of commitment to stay in the organization feel they have to stay because they estimate that the costs would be higher than the benefits resulting from leaving the organization [3, 11-13]. Often, an employer makes it harder to leave the organization by using many tricks and tools [13]. The predictors of commitment to stay in the organization are the personal characteristics of employees, alternatives and personal investments, while the benefits are reduced fluctuations in the organization, OCB, work performance and employee health. Commitment to stay in the organization can negatively influence attendance to work [13, 14].

3 Demographics, Job Characteristics and Characteristics of Organizations

Demographics is the study of general and particular population factors such as race, gender or occupation, as well as population density, size and location [15]. Demographics are the quantification of statistics for a given population and are used to identify the study of quantifiable sub-sets within a given population [16]. Demographic characteristics are widely used variables, in relation to organizational commitment and, as is shown in literature, there is a significant role of demographic factors in determining organizational commitment. Demographic factors such as age, gender, marital status, education and job tenure are included in many studies of the impact of demographic factors on commitment [e.g., 17]. Rabindarang, Bing and Yin [17] further emphasize that age is related to commitment in a way that older employees are more committed compared to younger employees and newcomers in an organization [see also, 18].

According to studies investigating the impact of gender on organizational commitment [e.g., 19], it is likely that age creates a feeling of organizational commitment depending on both experience and the conservative attitude it produces. Gender has a high impact on employees' organizational commitment, where it refers to socio-psychological categories of masculinity and femininity. As Pala et al. [19] further discuss, while some studies revealed that women are more committed to their organizations, other studies determined men as more committed than women. On the other hand, some other studies revealed that gender is unrelated to organizational commitment. It is also observed, that the cultural cluster or context can determine the impact of gender on organizational commitment in a way that if women are exposed to sex inequality in a certain context, it can affect their organizational commitment negatively [20]. Furthermore, it was found that there is the possible effect of gender on

organizational commitment can change and be affected, by an individual's hierarchical position and/or department within the organization [e.g., 19, 21].

Marital status is also a demographic factor, which influences commitment. Current literature shows that married people are more committed than single people. This is because they need a stable job, due to their perceived responsibility for their families [22]. It is clear that this commitment comes from concern for the economic safety of their families. Finally, education is yet another factor which can influence organizational commitment in a way that people with lower educational level and qualification are more committed to their organizations, as they rarely change their jobs. Conversely, Pala and colleagues [19] claim that there are studies, which reveal a direct positive relationship, between education level and an employees' commitment to their organization.

Many researchers have conducted studies to investigate the impact of various factors on organizational commitment, and in almost all of those studies, demographic factors were among those. For instance, Salami [23] found out in his study that all demographic factors except gender were significant predictors of organizational commitment among industrial workers. As he further states, some other researchers found out that education level and age were not significant predictors of organizational commitment, while others [e.g., 24] found significant relationship between job tenure and organizational commitment in their cultural context. Pala and colleagues [19] found in their research that years in occupation, gender, level of education and title, and meaning of the position in the organization were related to organizational commitment of health care workers in Turkey. Nifadkar and Dongre [25] found that age, gender, marital status and tenure are positively related, while, the level of education is negatively related to organizational commitment in India. Another study conducted in Pakistan showed that length of service is positively related, educational level is negatively related, while, age is not related to organizational commitment [26]. Amangala [16] found in his study that age, education, job position, and job tenure made significant impact on organizational commitment in Nigerian context. However, according to Salami [23], other researchers [27] found that demographic factors were not a significant predictor of organizational commitment.

The connection between job characteristics and employees' commitment is via motivation. Hackman and Oldham [28] argued that, in order to enhance employee motivation, every job must have five core characteristics: skill variety, task significance, task identity, autonomy and feedback. Job characteristics are, according to Greenberger and Strasser [29], the extent to which, a job is structured to provide regular feedback, as well as, a sense of task completion, and for employees to monitor their own behavior and gain an increased sense of personal control [30]. As they further argue, these job characteristics are attributes of job that motivate employees through the employees' perceived job characteristics, which further influence their motivation and determine their organizational commitment [31, 32]. Furthermore, perceived independence, sense of importance

and satisfaction with organizational demand all have significant impact on organizational commitment and so do specific characteristics of a job [33]. Many empirical studies [e.g., 27, 34, 35] provided evidence of strong correlations between dimensions of job characteristics and organizational commitment. On the other hand, some studies provided evidence of only a few dimensions of job characteristics significantly related to organizational commitment, dependent on the type of organization, type of job, position etc. In this paper, the authors argue that perceived organizational characteristics also have impact on organizational commitment.

4 Research Methodology

The main question, to which this paper is intended to give explicit answer, is whether demographic characteristics of employees and the characteristics of their jobs and organizations influence their organizational commitment in Central European transitional and post-transitional context? When taking this question into account, three basic hypotheses were created, representing the perceived research problem:

- **H1** Employees with different demographic characteristics manifest different levels of organizational commitment
- **H2** Different characteristics of jobs induce different level of employee organizational commitment;
- **H3** Various characteristics of organizations induce different level of employee organizational commitment.

4.1 Sample and Procedure

Research for this study was conducted between 2012 and 2014. It took place in various private and public (state-owned) organizations, from various fields of operation and various organizational sizes from Serbia's Central European Autonomous Province of Vojvodina, as well as, Hungary's Csongrád and Bács-Kiskun Counties as a part of a broader study, which in its focus included investigations of the relationships between various leadership and organizational behavior variables. Results of this study, as well as, the theoretical and methodological background were published in 2015, in papers and conference proceedings [e.g., 36, 37]; the overall results are planned to be published during 2015/2016. The data collection method included both physically distributed questionnaires and online surveys, however, the first method being the dominant one. A total of 1400 questionnaires were distributed with a return rate of 855 or a 61% return rate. Along with the online survey, the total number is 891.

4.2 Description of the Sample

Females make 62%, while males 38% of the respondents. 16 respondents did not specify their gender. The mean for the age is 40.526 years (SD = 11.030). 28.2% of the respondents belong to 30-39 years of age category, 26.6% were 50-59 years, 22.4% were 40-49 years, 20.2% were 20-29 years, 2.3% were over 60 years and only 0.4% were less than 19 years old. Regarding tenure with the organization, the mean is 12.149 years (SD = 10.659). 50% of the respondents worked 0-9 years in the current organization, 23.5% 10-19 years, 16.4% 20-29 years, and 10.1% 30-39 years. 23 respondents did not give answer to this question. As for the total years of service (total tenure) of the respondents, the mean is 14.595 years (SD = 11.120). 42.3% of the respondents had 0-9 years of service, 23% had 10-19 years of service, 21.4% had 20-29 years of service, 12.3% had 30-39 years of service, and only 1% over 40 years of service (with the maximum of 44 years of service). 357 respondents did not give an answer to this question due to organizational policies. There were eight levels of education offered in the questionnaire, with the results: primary school (3.1%), secondary school (50.6%), college (16.4%), bachelor's degree (18%), master's degree (7.5%), medical degrees (M.D. and specialists; 4.1%), Ph.D. (0.2%), and other (0.2%). 5 respondents did not specify their education.

Distribution of respondents by job type: 27.4% was performing jobs related to simple task execution (operational jobs), 17.3% technical jobs, 16.9% jobs related to education, 14.3% administrative jobs, 9.5% jobs that require high level of expertise, 7.4% management, 5% jobs related to communication with stakeholders, and 2.2% other jobs. 356 respondents did not specify their job type due to organizational policies. Hierarchical position in the organization: 82.4% of the respondents were not on a managerial position. 360 respondents did not specify their position status due to organizational policies.

Distribution of the organizations the respondents work in by ownership: 73.3% state-owned (public) organizations, 25.2% private organizations, 0.6% non-governmental organizations, and 0.9% unknown. Distribution of the organizations the respondents work in according to their dominant activity: 59.8% public service organizations, 20.3% manufacturing organizations, 15.8% service organizations, 2.3% manufacturing and service organizations, 0.6% non-governmental organizations, 0.8% other types of organizations, and 0.4% unknown.

4.3 Instruments

Two questionnaires were used in this part of the research. The first questionnaire is a general one, intended to gather basic information about the respondents, their job and organization, this questionnaire, was created by the authors for use in this study. The goal was to collect data on general demographic characteristics of the

respondents, nature of their job, hierarchical position they hold in the organization, as well as general information about the organization they work in. Some of the elements of this questionnaire are age, education, job type, position, tenure, type of organization, etc.

The second questionnaire measured Organizational Commitment with the widely used 15-item Organizational Commitment Questionnaire - OCQ [4], on a standard 5-point Likert Scale, from "completely disagree" to "completely agree". This questionnaire is the most widely used instrument for measuring organizational commitment [c.f., 14, 36, 38], with investigated and proven psychometric characteristics and is often used, for measuring commitment within a wide range of job categories [39]. It includes items concerning the employee's perceptions about their loyalty to the organization, their willingness to completely, engage in activities that achieve organizational aims and their acceptance of organizational values [4].

Item number 7 of the questionnaire, was statistically problematic, since it has a very low component saturation, accordingly this item was eliminated from further analysis. Cronbach's alpha is very high, the improved questionnaire has high reliability (α = .901). The representativeness of the items according to the KMO criterion is very significant (.930). In further analyses of the main scale, the items with reversed directions, specifically, items 3, 9, 11, 12 and 15, were recoded. This questionnaire offers the possibility to create two subscales in the results analysis [see also, 40]. The first subscale (Cronbach's alpha = .909, Λ=6.427, includes 45.908% of the total variance) refers to the respondents value commitment, which reflects their affective commitment and includes items 1, 2, 4, 5, 6, 8, 10, 13, and 14. The second subscale (Cronbach's alpha = .739, Λ=1.450, includes 10.356% of the total variance) refers to the respondents commitment to stay, which reflects their continuance commitment and includes items 3, 9, 11, 12, and 15.

4.4 Data Processing Methods

The data in this research were completely analyzed using IBM SPSS statistics software. Analyzes included instrument checks (Cronbach's alpha, Guttman-Kaiser, factor analyses, representativeness, validity, Cattel's SCREE test), analysis of the distribution of scores, descriptive statistics for scores (Mean, SD, Skewness, Kurtosis, Kolmogorov-Smirnov), correlations (Spearman's), t-test, and Kruskal-Wallis test.

5 Results

5.1 Gender Differences

The analysis of differences in commitment levels did not reveal any significant variances among different genders for any of the variables (Table 1).

5.2 Relations to Age, Tenure, Service Years and Education

The variations in commitment level resulting from differences in age, education, tenure in current organization, and total service years of the respondents were determined with Spearman's rank correlations (Table 2). Commitment to the values of the organization is positively related with total years of service (ρS=.188, $p \leq 0.01$), age (ρS=.099, $p \leq 0.01$) and education (ρS=.096, $p \leq 0.01$). Commitment to stay with the organization is positively related with education (ρS=.204, $p \leq 0.01$) and age (ρS=.203, $p \leq 0.01$), and negatively related with tenure in current organization (ρS=-.090, $p \leq 0.01$). Overall, organizational commitment is positively related only with total years of service (ρS=.178, $p \leq 0.01$).

Table 1
t test results for gender differences

	Levene's test		t test for independent samples						
	F	p	t	df	p	Group	N	M	s
Commitment to organizational values	3.135	.077	-.680	835	.497	Males	318	31.7829	8.2412
						Females	519	32.1631	7.6058
Commitment to stay with the organization	.650	.420	.279	835	.780	Males	318	17.3969	4.3571
						Females	519	17.3125	4.1778
Organizational commitment	.193	.660	-.792	835	.428	Males	318	44.3792	6.5427
						Females	519	44.7419	6.3569

Table 2
Spearman's rank correlations for age, education, tenure, and years of service[1]

		Age	Education	Tenure in current org.	Total service years
Commitment to organizational values	ρ_S	.099**	.096**	-.017	.188**
	p	.004	.005	.618	.000
	N	848	846	830	496

[1] Legend: ** Significance at $p \leq 0,01$ level

Commitment to stay with the organization	ρ_S	.203**	.204**	-.090**	.035
	p	.000	.000	.009	.432
	N	848	846	830	496
Organizational commitment	ρ_S	.020	.020	.025	.178**
	p	.561	.555	.465	.000
	N	848	846	830	496

5.3 Differences Related to Position in the Organization

The results of the t-test revealed that there is a statistically significant difference between respondents in managing and non-managing positions within scores of commitment, to the values of the organization ($t=3.852$, $p\leq0.01$) and commitment to stay with the organization ($t=4.107$, $p\leq0.01$), with higher levels of commitment to the values of the organization and commitment to stay with the organization among respondents on managing positions (Table 3).

Table 3

t test results for differences related to position

	Levene's test		t test for independent samples						
	F	p	t	df	p	Group	N	M	s
Commitment to organizational values	5.480	0.020	3.852	149.656	0.000	Managing	87	35.506	6.121
						Non-manag.	406	32.586	7.650
Commitment to stay with the organization	0.883	0.348	4.107	491	0.000	Managing	87	19.421	4.149
						Non-manag.	406	17.355	4.282
Organizational commitment	0.006	0.938	1.202	491	0.230	Managing	87	45.944	5.943
						Non-manag.	406	45.116	5.808

5.4 Differences Related to Job Type

Differences in scores related to job type were determined with the help of nonparametric Kruskal Wallis test. The results revealed differences in level of commitment to the values of the organization and commitment to stay with the organization (Table 4).

Table 3

Results of nonparametric Kruskal Wallis test for differences in scores related to job type

	Hi square	df	p
Commitment to organizational values	15.704	7	.028
Commitment to stay with the organization	39.427	7	.000
Organizational commitment	6.425	7	.491

Further tests revealed that respondents performing educational and management jobs have highest scores, while respondents performing operational jobs have lowest scores of commitment to the values of the organization. Also, respondents performing educational, high expertise and management jobs have the highest scores, while respondents performing operational jobs have lowest scores of commitment to stay with the organization.

5.5 Differences Related to Ownership over the Organization

Ownership of the organization has little impact on commitment. The results of the t test revealed that there is statistically less significant difference between respondents from state-owned and private organizations only in commitment to stay with the organization scores ($t=-2.343$, $p\leq0.05$), with higher level of commitment to stay in private organizations (Table 5).

Table 5

t test results for differences related to ownership over the organization

	Levene's test		t test for independent samples						
	F	p	t	df	p	Group	N	M	S
Commitment to organizational values	3.304	.069	-.970	838	.332	State-owned	625	31.803	8.056
						Private	215	32.405	7.221
Commitment to stay with the organization	.870	.351	-2.343	838	.019	State-owned	625	17.094	4.269
						Private	215	17.880	4.160
Organizational commitment	6.556	.011	.108	469.510	.914	State-owned	625	44.591	6.737
						Private	215	44.542	5.288

5.6 Differences Related to the Dominant Activity of the Organizations

Differences in commitment related to the dominant activity of the organization were analyzed with, one-way analysis of variance (ANOVA, Table 6). The results (Table 8) showed that there is statistically significant difference between subgroups of respondents in different types of organizations in commitment to the values of the organization ($F(2;815) = 4.197$, $p\leq0.05$), as well as in commitment to stay with the organization ($F(2;815) = 3.988$, $p\leq0.05$).

Post-hoc test – least significant difference (LSD) has shown that respondents employed in manufacturing organizations have considerably higher score in commitment to the values of the organization from respondents employed in a public service and service organizations (Table 7).

Post-hoc test – least significant difference (LSD) has shown that respondents employed in manufacturing organizations have considerably higher score in

commitment to stay with the organization from respondents employed in public service organizations (Table 8).

Table 6
One-way analysis of variance – ANOVA

		Sum of squares	df	Square middle	F	p	Levene's statistics	P
Commitment to organizational values	Between groups	521.223	2	260.612	4.197	.015	.479	.620
	Inside groups	50606.443	815	62.094				
	Total	51127.666	817					
Commitment to stay with the organization	Between groups	143.883	2	71.941	3.988	.019	.016	.984
	Inside groups	14703.993	815	18.042				
	Total	14847.876	817					
Organizational commitment	Between groups	126.333	2	63.167	1.552	.212	1.859	.156
	Inside groups	33168.662	815	40.698				
	Total	33294.995	817					

Table 7
Post-hoc test – LSD for commitment to organizational values

I	J	Difference M (I-J)	Std. error	P
Public-service	Manufacturing	-1.93602*	.69331	.005
	Service	.08283	.76270	.914
Manufacturing	Service	2.01884*	.90492	.026

Table 8
Post-hoc test – LSD for commitment to stay with the organization

I	J	Difference M (I-J)	Std. error	P
Public-service	Manufacturing	-1.05448*	.37372	.005
	Service	.22052	.41112	.592
Manufacturing	Service	-.83396	.48778	.088

6 Discussion

6.1 Influence of the Demographic Characteristics of Employees, on Their Commitment Levels

The results did not reveal significant differences among males and females for any of the variables. However, certain differences related to age, education, tenure in the current organization, and total years of service do exist, meaning that gender does not have influence, while age, education, tenure, and years of service have

influence or commitment. This finding are not consistent with the results of previous studies conducted in western developed environments, which mostly indicate that women are getting better, more demanding and accountable jobs, which results in higher satisfaction and commitment to the organization. These studies also indicate that women as their future goals states increase in commitment, while man are more prone to refer their expectations and change of job.

Employees with higher number of total years of service, higher education and older employees demonstrated higher levels of commitment to the values of the organization. Reasons for this are experience that comes with years of service, age and education that enables employees to better understand and adopt the values of the organization as well as to harmonize those values with their own values and goals. Further, employees with higher education and professional qualifications often have stronger ambitions and desire for advancement, so they are committed to the values of the organization in order to achieve organizational goals, therefore achieving their own goals as well. On the other hand, employees with lower number of total years of service, lower education as well as younger employees, often have too high expectations and are holding on to their personal ideals that, in some situations, prevents them to realistically perceive and evaluate the values and goals of the organization.

Higher commitment to stay within the organization belongs to older employees, employees that have a shorter job tenure in the current organization and employees with higher education. This result, directly, reflects the specificities of Central European economy and labor market. In highly developed countries, individuals with better education have far more opportunities for employment, so fluctuations, i.e. leaving of well-educated personnel to better and more successful organizations is not a rare phenomenon, therefore, their commitment to stay with the current organization is significantly lower. However, the picture in Central Europe is very different, where the employment possibilities of highly educated human resources are considerably worse. Therefore, the commitment to stay with the current organization with highly educated personnel is higher, since they are well aware of the fact that they do not have much choice. Further, older employees are well aware that for them it is especially difficult to find another job, while employees that have a shorter job tenure in the organization, i.e. employees that had been recently employed, are satisfied that they managed to get a job at all, often after long period of job seeking. This fact drives their commitment to stay within the organization, while employees that are longer in the organization have a desire to change to better jobs and organizations.

Employees with longer service are the only ones who have higher overall organizational commitment, but that does not mean that other employees are not committed to their organizations. It means just that employees with longer service have slightly higher commitment, i.e. they are the most committed employees to their organizations.

6.2 Influence of Job Characteristics on Employee Commitment Levels

The influence of job characteristics is determined with regard to two categories: position in the organizational hierarchy (managing/non-managing position) and job type (jobs related to simple task execution – operational jobs, technical jobs, jobs related to education, administrative jobs, jobs that require high level of expertise, management jobs, jobs related to communication with stakeholders and other jobs).

The results of the analysis showed that there is a difference in commitment to the values of the organization and commitment to stay within the organization among managing and non-managing groups of employees, with significantly better results of employees on managing positions. Better results of employees on these positions were definitely expected. Higher position in the organizational hierarchy can bring numerous benefits to employees, ultimately leading to higher commitment of those employees. Employees in managing positions belong to an often called "inner group" of employees, having access to strategic resources, better relations with the executives of the organization and higher levels of accountability, leading to high commitment, satisfaction and performance [41, 42, 43].

In the second category, a difference was discovered, among employees conducting different job types in their commitment to the values of the organization and their commitment to stay with the organization. Table 9 shows the best and worst ranked job types. Overall, employees performing managerial and education related jobs have highest scores, while employees conducting operational jobs (jobs related to simple task execution) have lowest scores on both variables.

Table 9

Best and worst ranked job types overview

Commitment type	Best ranked jobs	Worst ranked jobs
Commitment to stay with the organization	• Education related jobs • High expertise jobs • Managerial jobs	• Other types of jobs • Operational jobs
Commitment to the values of the organization	• Education related jobs • Managerial jobs	• Operational jobs

6.3 Influence of the Characteristics of Organizations on Employee Commitment Levels

The characteristics of organizations were studied with regard to two groups of characteristics: organization type by dominant activity and ownership over the organization. The analysis revealed that the characteristics of organizations have limited influence. Organization type have limited influence on commitment to the values of the organization and commitment to stay within the organization, while ownership over the organization have very limited effect only on commitment to stay within the organization. Employees in manufacturing organizations are much more committed to the values of the organization from employees in public service and service organizations. Employees in manufacturing organizations are also much more committed to stay within the organization from employees in public-service organizations. Further, employees in privately owned organizations are more committed to stay within the organization from employees in state owned organizations, but with a low significance of the difference between them.

Many studies in the past had been engaged in researching the commitment of employees in public organizations. Those studies had mostly similar conclusions related to public servants having higher level of commitment to stay within the organization from employees in other organizations [13, 44]. In those settings, job security and employee ethics had the most influence on high commitment of public servants [44, 45]. However, the results of this study revealed opposite results, with employees in privately owned and manufacturing organizations being more committed. The reasons for the different results are only a guess; whether it is because previous studies focused only on western highly developed environments or simply time has changed.

Conclusion

The main question of this paper, was supposed to provide an answer to whether or not the demographic characteristics of employees and the characteristics of their jobs and organizations influence their organizational commitment in a Central European settings? In general, several main conclusions are drawn and from the results of the analysis:

- Gender does not have any influence on organizational commitment
- Characteristics of organizations and most demographic characteristics have little effects on organizational commitment
- Job characteristics have a strong impact on commitment.

Based on the results, hypotheses H1, assuming that demographic characteristics of employees influence the level of their organizational commitment is partially supported, since gender of the employees did not have any influence, while other characteristics did have influence, such as, tenure with the current organization,

formal education, professional qualifications and age influence commitment to stay with the organization. Furthermore, total years of service, formal education and age influence commitment to the values of the organization. Total years of service, is the only characteristic that overall influences organizational commitment. While not all personal characteristics affect every variable equally, it can be concluded that, personal characteristics of employees influence the level of their commitment. Hypotheses H2, assuming that job characteristics influence the level of employee organizational commitment is completely supported, since both job type and hierarchical position have significant influence on organizational commitment. Hypotheses H3, assuming that the characteristics of organizations influence the level of employee organizational commitment is only partially supported, since the limited influence of both groups of the following characteristics: organization type and ownership over the organization. Organization type has a limited influence on commitment to the values of the organization and commitment to stay within the organization, while ownership over the organization has a very limited effect on the commitment to stay with the organization.

Acknowledgements

This research was supported by the Provincial Secretariat for Science and Technological Development, of the Autonomous Province, of Vojvodina, as part of, the project "Effects of Organizational Communication on Employee Organizational Behavior". The funding was granted to Dr. Valentin Kónya, the project leader.

References

[1] J. M. George and G. R. Jones: Understanding and Managing Organizational Behavior. Pearson Education Limited, 2011

[2] P. C. Morrow and J. C. McElroy: Introduction: Understanding and Managing Loyalty in a Multi-Commitment World, Journal of Business Research, Vol. 26, 1993, pp. 1-2

[3] J. P. Meyer, N. J. Allen, and L. Topolnytsky: Commitment in a Changing World of Work, Canadian Psychology-Psychologie Canadienne, Vol. 39, 1998, pp. 83-93

[4] L. W. Porter, R. M. Steers, R. T. Mowday, and P. V. Boulian: Organizational Commitment, Job Satisfaction, and Turnover among Psychiatric Technicians, Journal of Applied Psychology, Vol. 59, 1974, pp. 603-609

[5] B. Buchanan: Building Organizational Commitment: The Socialization of Managers in Work Organizations, Administrative Science Quarterly, Vol. 19, 1974, pp. 533-546

[6] G. Johns and A. M. Saks: Organizational Behaviour: Understanding and Managing Life at Work, 6 ed. Canada: Pearson Education / Prentice Hall, 2005

[7] J. R. Schermerhorn, J. G. Hunt, R. N. Osborn, and M. Uhl-Bien: Organizational Behavior, 11 ed. Hoboken, NJ: John Wiley & Sons, 2010

[8] R. M. Kanter: Commitment and Social Organization: A Study of Commitment Mechanisms in Utopian Communities, American Sociological Review, Vol. 33, 1968, pp. 499-517

[9] T. S. Bateman and S. Strasser: A Longitudinal Analysis of the Antecedents of Organizational Commitment, The Academy of Management Journal, Vol. 27, 1984, pp. 95-112

[10] C. E. Rusbult and D. Farrell: A Longitudinal Test of the Investment Model: The Impact on Job Satisfaction, Job Commitment, and Turnover of Variations in Rewards, Costs, Alternatives, and Investments, Journal of Applied Psychology, Vol. 68, 1983, pp. 429-438

[11] N. J. Allen and J. P. Meyer: The Measurement and Antecedents of Affective, Continuance and Normative Commitment to the Organization, Journal of Occupational Psychology, Vol. 63, 1990, pp. 1-18

[12] N. J. Allen and J. P. Meyer: Affective, Continuance, and Normative Commitment to the Organization: An Examination of Construct Validity, Journal of Vocational Behavior, Vol. 49, 1996, pp. 252-276

[13] J. P. Meyer and N. J. Allen: Commitment in the Workplace: Theory, Research, and Application. Thousand Oaks, CA: SAGE, 1997

[14] J. P. Meyer, D. J. Stanley, L. Herscovitch, and L. Topolnytsky: Affective, Continuance, and Normative Commitment to the Organization: A Meta-Analysis of Antecedents, Correlates, and Consequences, Journal of Vocational Behavior, Vol. 61, 2002, pp. 20-52

[15] J. Blythe: Essentials of Marketing. Harlow, Essex, England: Pearson Education Limited, 2005

[16] T. A. Amangala: The Effect of Demographic Characteristics on Organizational Commitment: A Study of Salespersons in the Soft Drink Industry in Nigeria, European Journal of Business and Management, Vol. 5, 2013, pp. 109-118

[17] S. Rabindarang, K. W. Bing, and K. Y. Yin: The Impact of Demographic Factors on Organizational Commitment in Technical and Vocational Education, Malaysian Journal of Research, Vol. 2, 2014, pp. 56-61

[18] H. Hulpia, G. Devos, and Y. Rosseel: Development and Validation of Scores on the Distributed Leadership Inventory, Educational and Psychological Measurement, Vol. 69, 2009, pp. 1013-1034

[19] F. Pala, S. Eker, and M. Eker: The Effects of Demographic Characteristics on Organizational Commitment and Job Satisfaction: An Empirical Study on Turkish Health Care Staff, The Journal of Industrial Relations and Human Resources, Vol. 10, 2008, pp. 54-75

[20] H. Y. Ngo, A. Wing, and N. Tsang: Employment Practices and Organizational Commitment: Differential Effects for Men and Women?, The International Journal of Organizational Analysis, Vol. 6, 1998, pp. 251-266

[21] A. Cohen: Antecedents of Organizational Commitment across Occupational Groups: A Meta-Analysis, Journal of Organizational Behavior, Vol. 13, 1992, pp. 539-558

[22] Y. O. Choong, C. E. Tan, C. G. Keh, Y. H. Lim, and Y. T. Tan: How Demographic Factors Impact Organisational Commitment of Academic Staffs in Malaysian Private Universities: A Review and Research Agenda, International Journal of Academic Research, Vol. 4, 2012, pp. 72-76

[23] S. O. Salami: Demographic and Psychological Factors Predicting Organizational Commitment among Industrial Workers, Anthropologist, Vol. 10, 2008, pp. 31-38

[24] S. de los Santos and E. Not-Land: Factors Related to Commitment of Extension Professionals in the Dominican Republic: Implications for Theory and Practice, Journal of Agricultural Education, Vol. 35, 2006, pp. 57-63

[25] R. S. Nifadkar and A. P. Dongre: To Study the Impact of Job Satisfaction and Demographic Factors on Organizational Commitment among Girls' College, Pune, India, Journal of Business Management & Social Sciences Research, Vol. 3, 2014, pp. 1-8

[26] A. Iqbal: An Empirical Assessment of Demographic Factors, Organizational Ranks and Organizational Commitment, International Journal of Business and Management, Vol. 5, 2010, pp. 16-27

[27] J. E. Mathieu and D. M. Zajac: Review and a Meta-Analysis of the Antecedents, Correlates and Consequences of Organizational Commitment, Psychological Bulletin, Vol. 108, 1990, pp. 171-194

[28] J. R. Hackman and G. R. Oldham: Motivation through the Design of Work: Test of a Theory, Organizational Behavior and Human Performance, Vol. 16, 1976, pp. 250-279

[29] D. B. Greenberger and S. Strasser: The Development and Application of a Model of Personal Control in Organizations, Academic of Management Review, Vol. 11, 1986, pp. 164-177

[30] H. Obi-Nwosu, J. A. O. Chiamaka, and O. M. Tochukwu: Job Characteristics as Predictors of Organizational Commitment among Private

Sector Workers in Anambra State, Nigeria, International Journal of Asian Social Science, Vol. 3, 2013, pp. 482-491

[31] S. F. Chiu and H. L. Chen: Relationship between Job Characteristic and Organizational Citizenship Behaviour: The Mediational Role of Job Satisfaction, Social Behaviour and Personality, Vol. 33, 2005, pp. 523-540

[32] C. J. Mottaz: Determinants of Organizational Commitment, Human Relations, Vol. 41, 1988, pp. 467-482

[33] I. E. Jernigan and J. M. Beggs: An Examination of Satisfaction with my Supervisor and Organizational Commitment, Journal of Applied Social Psychology, Vol. 35, 2005, pp. 2171-2192

[34] N. T. Feather and K. A. Rauter: Organizational Citizenship Behaviours in Relation to Job Status, Job Insecurity, Organizational Commitment and Identification, Job Satisfaction and Work Values, Journal of Occupational and Organizational Psychology, Vol. 77, 2004, pp. 81-94

[35] R. I. Allen, E. G. Lambert, S. Pasupuleti, T. Cluse-Tolar, and L. A. Ventura: The Impact of Characteristics on Social and Human Service Workers, Social Work and Society, Vol. 2, 2004, pp. 173-188

[36] V. Kónya, L. Grubić-Nešić, and D. Matić: The Influence of Leader-member Communication on Organizational Commitment in a Central European Hospital, Acta Polytechnica Hungarica, 2015, Vol. 17, No. 12, pp. 109-128

[37] V. Kónya, D. Matić, J. Pavlović: Unconventional Approaches in Leadership Research, Proceedings of the 4th International scientific-professional conference "PAR International Leadership Conference: Change Leadership – Key to Successful Growth (PILC 2015)", March 13-14, Business School PAR Rijeka, In press

[38] J. P. Meyer and N. J. Allen: Testing the 'Side-Bet Theory' of Organizational Commitment: Some Methodological Considerations, Journal of Applied Psychology, Vol. 69, 1984, pp. 372-378

[39] R. T. Mowday, R. M. Steers, and L. W. Porter: The Measurement of Organizational Commitment, Journal of Vocational Behavior, Vol. 14, 1979, pp. 224-247

[40] H. L. Angle and J. L. Perry: An Empirical Assessment of Organizational Commitment and Organizational Effectiveness, Administrative Science Quarterly, Vol. 26, 1981, pp. 1-14

[41] G. Graen and M. Uhl-Bien: Relationship-based Approach to Leadership: Development of Leader-Member Exchange (LMX) Theory of Leadership over 25 Years: Applying a Multi-Level Multi-Domain Perspective, The Leadership Quarterly, Vol. 6, 1995, pp. 219-247

[42] M. D. Zalesny and G. Graen: Exchange Theory in Leadership Research, in Encyclopedia of Leadership, G. Reber, Ed. Linz: Linz University Press, 1986

[43] H. J. Klein and J. S. Kim: A Field Study of the Influence of Situational Constraints Leader-Member Exchange, and Goal Commitment on Performance, Academy of Management Journal, Vol. 41, 1998, pp. 88-95

[44] J. L. Perry, Antecedents of Public Service Motivation: Journal of Public Administration Research and Theory, Vol. 7, 1997, pp. 181-197

[45] K. T. Liou: Professional Orientation and Organizational Commitment Among Public Employees: An Empirical Study of Detention Workers, Journal of Public Administration Research and Theory, Vol. 5, 1995, pp. 231-246

An Overview of Low-Cost EGSE Architectures Improvement

Sándor Szalai[1], János Nagy[1], István Horváth[1], Bálint Sódor[1], Gábor Tróznai[1], Kálmán Balajthy[2], János Sulyán[2]

[1]Wigner Research Center for Physics, Konkoly-Thege u. 29-33, 1125 Budapest, Hungary, E-mail: {szalai.sandor, nagy.janos, horvath.istvan, sodor.balint, troznai.gabor}@wigner.mta.hu

[2]SGF Ltd., Pipiske u. 1-5/20, 1125 Budapest, Hungary; {balajthy, sulyan}@sgf.hu

Abstract: this article presents EGSE architecture improvement from the 1980s until nowadays following hardware development. In EGSE development we looked for cost-effective solutions by applying available industrial products. We started EGSE development for VEGA-Halley mission in the 1980s, based on a microprocessor standalone system. The next stages of EGSE architecture were based on IBM compatible PCs with dedicated interface cards. The subsequent generation of EGSE consisted of two physical units, one was a commercial computer and the other one was an embedded processor card for signal level simulation. There was a serial communication line between the units. The fourth generation of EGSE contains high speed bus for internal communication and the use of embedded processor made simulation and data acquisition possible in real-time. The software was developed in assembly in the first generation of EGSE. The further operating software runs on a distributed intelligence system containing Windows and real-time Linux platforms.

Keywords: EGSE; VEGA-Halley; Spectrum-X-Ray-Gamma; Rosetta; Comet; Churyumov–Gerasimenko; lander

1 Introduction

The task of EGSE (Electrical Ground Support Equipment) is to support the development and test of flight units. The EGSE supports all phases of assembly, integration and final validation test. In this paper, we present improvements in EGSE architecture development over the past 30 years. The subsequent EGSE architecture for missions followed hardware performance improvement, and software technology also followed the improvements provided by the hardware. During these three decades we worked with four different EGSE generations. In the 1980s we started EGSE development for the VEGA-Halley mission, based on

a microprocessor standalone system. In the next stages of EGSE architecture to reduce development costs they were based on commercially available computers – typically IBM compatible PCs – extended with dedicated interface cards, which used the resources of standardized computers. The next generation of EGSE consisted of two physical units, one was a commercial computer, the other one was a signal level simulator controlled by an embedded processor. This latter one contained either dedicated interface cards – partly self-developed – or widely used industry-standard cards – for signal level, a simulation standardised serial communication line was used between the units. The fourth generation of EGSE contains high-speed bus (Ethernet) for internal communication. The use of an embedded processor made simulation and data acquisition possible in real-time.

The software work started during assembly in the first generation of EGSE. Current operating software runs on a distributed intelligence system containing Windows and real-time Linux platforms. Running Windows on a commercial computer offers the advantages of user-friendly interface of Graphical User Interface (GUI) based on LabWindows or Java, efficient data storage, and processing capability. A wide range of graphic software development tools is available for Windows, which helps the fast and efficient development of GUI. Linux, extended with real-time facilities allows for the running of real time simulation and data acquisition on the embedded processor. The software environment insures a lot of advantages: the user can control all functions through GUI, definition timed sequence of commands, decoding and visibility of housekeeping packets, mathematical operation can be performed on data, e.g. polynomial interpolation, Fourier transformation, etc. Commands can be contained in a macro file with pre-written timings. The housekeeping data can be displayed in user- friendly form. The conversion is controlled by a simple structured file, which can be easy modified even by a non-skilled user.

Depending on actual space-probe and onboard instruments some functions can be left out (Sensor Stimulator) or can be multiplied (Fast and Slow telemetry or SpaceWire and Mil1553).

In this article, we present how EGSE which has been developed for our projects to test space instruments has changed since the early 1980s till nowadays as a result of technical development.

Figure 1 simplified functional block diagram of EGSE.

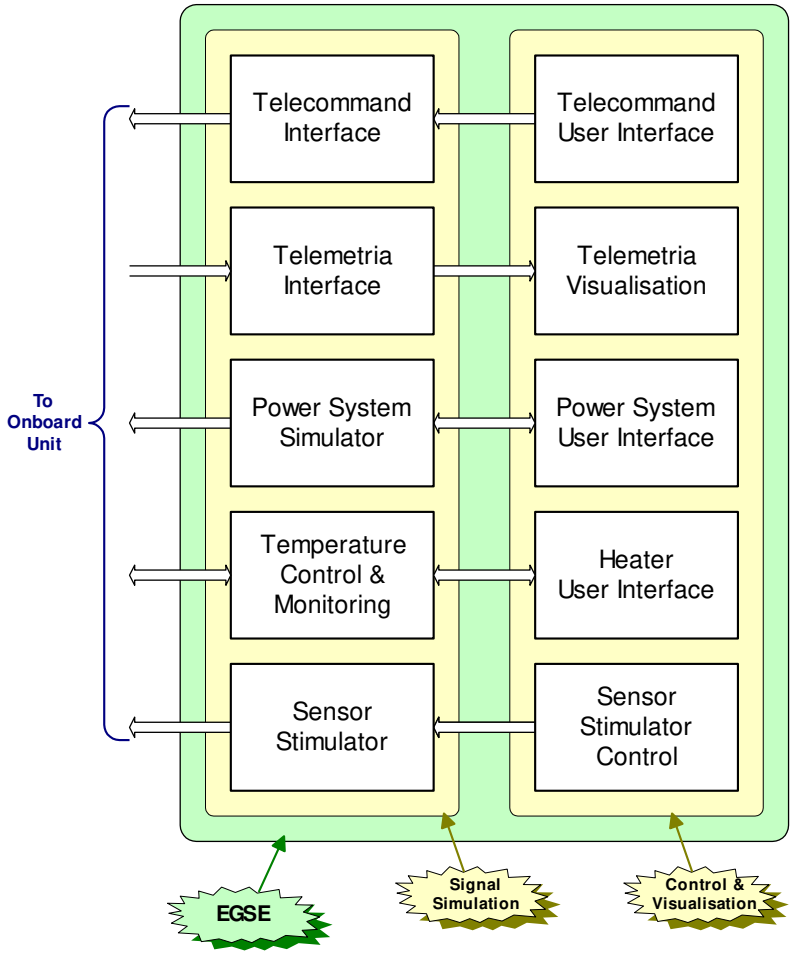


Figure 2
Simplified functional block diagram of EGSE

2 EGSE from the 1980s

Our institute joined the experiment Halley-VEGA in the 1980s: we took part in television system development [1, 2, 3]. The mission was aimed at investigating and observing the comet Halley, and to broadcast pictures of the comet during its approach. Key dates of the project include launch December 1984 and flyby March 1986. The mission ensured double redundancy by doubling the probe. The probes were called VEGA1 and VEGA2. The distance of the nearest approach

was 8,900 km by VEGA1 and 8,030 km by VEGA2. Our institute designed and built several instruments for the VEGA mission, e.g. the electronics of imaging and tracking system, the so-called TV system. The onboard television system controlled the approach phase in the near region of the comet. During this phase the transmission time of commands between the space probe and the ground control center required too much time in comparison with flyby times. The close approach duration was three hours. It was the first time in the history of space exploration when autonomous control was based on a real-time image processing. For testing the tracking system (hardware and software) we had to realize not only usual EGSE, but a special tracking loop including optical parts to simulate the accurate movement (relative orbit) of comet nucleus in the field of view of the TV system.

This time different autonomous simulators were used to test the onboard equipment. Embedded 8 bit microprocessor based systems were developed to test every single research equipment. The control of tested instruments was realized by knobs and switches; the interpretation of telemetry information occurred on indicator lamps.

The test equipment had different jobs which ensured testing comet recognition using hardware and software as well as checking and calibration of the onboard system. The features of the test system are aimed at simulating operation circumstances. The EGSE of TV system (the Russian abbreviation of EGSE is KIA) was based on a microprocessor system. The core of test equipment was a Z80 processor with individual UMDS bus (Universal Microprocessor Development System) developed by our Institute. It was able to test and control all interfaces of the TV system. A second microprocessor generated the nucleus orbit images for testing the tracking accuracy. The operation system Z80-RIO (Re-locatable Modules and I/O Management) was also individually developed. It enabled monitoring and debugging functions. The structured software contained elements to test either individual units of the TV-System or perform complex tests of several units working together or the entire system's combined cooperation. The embedded software was developed in assembly language.

Figure 3
The autonomous test equipment for the Tünde instrument of VEGA-Halley mission, the top box of the EGSE is the Power Supply Simulator

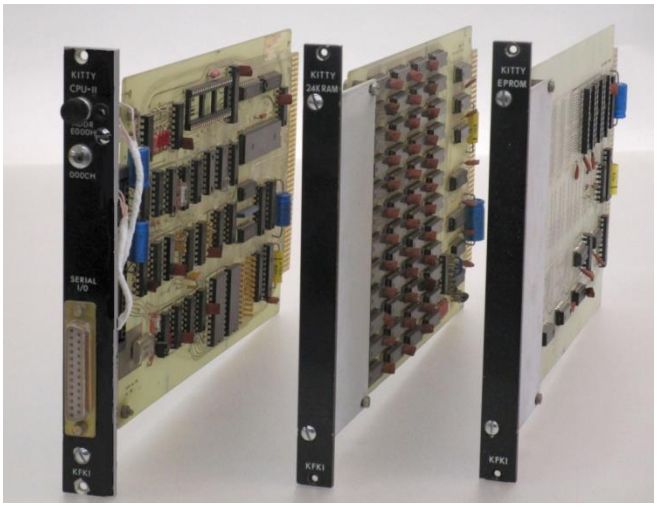


Figure 4
Test cards for VEGA experiment

3 EGSE in the First Half of the Nineties

The project was an international astrophysical project. The space probe was originally planned to be launched in 1998. However, the mission was cancelled after a ten-year delay. The scientific objects of the project included observation of known, as well as discovery of new gamma sources. In our institute the onboard data acquisition computer BIUS and other instruments and their EGSE were prepared. Besides Spectrum-X-Ray-Gamma, in the nineties we took part in other missions where similar structures of EGSE were applied. These missions were MARS96 and Cassini. The dedicated interfaces which simulated a certain electric surface of spacecraft was connected to the standardized PC bus (so-called ISA bus). Through this ISA bus the simulator units could use resources of the PC. The telemetry simulator interfaces used the PC memories through Direct Memory Access (DMA).

Basic activities of BIUS are the following:

- collecting scientific and technical data from the experiments and storing them in the on-board storage memory,

- preprocessing science data and sending them to the Earth over radio link,

- controlling scientific experiments according to a predefined cyclogram or the uplinked Earth commands.

As a result of the spread of PCs in the early nineties, PCs were introduced to and applied in many fields. Instead of processor unit development, PCs offered a quick alternative to the realization and implementation of control, management and data collection with computers.

The electrical ground support equipment is an IBM-PC based test system, in which special interface cards simulate the space probe's electrical signals. The system includes the following individually developed circuit cards:

1. On-board data acquisition and control bus simulator;
2. Analogue telemetry and relay command simulator;
3. Coded command and on-board time;
4. Fast telemetry simulator;
5. Slow telemetry simulator;
6. Inner bus simulator ("processor bus").

The operating software of the EGSE was written partly in Borland C++. The software is menu-driven, window oriented, quasi-real time and interactive. The control commands are generated through the keyboard and forwarded to the Coded-Command and time simulator card by programmed (polled) method. It sends them by hardware method to the on-board system. Receiving fast telemetry signals is organized as a background job (direct memory access).

There are more advantages of using a PC in EGSE: possibility of applying industrial cards for interface testing, well defined bus system and great volume of software support on the widely available hardware structure, which supported software development, evaluation and data acquisition.

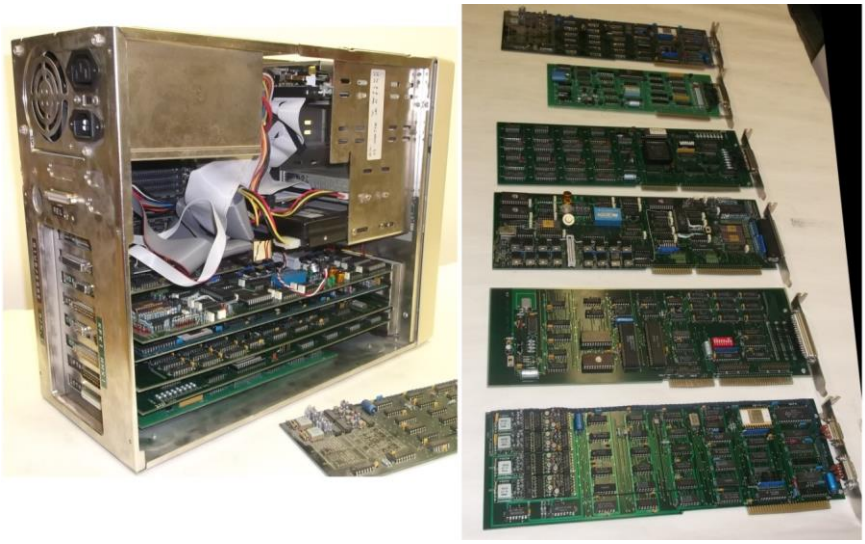


Figure 5
EGSE of Spectrum-X-Ray-Gamma EGSE. It was based on a commercial PC with industry made and self-developed extension cards.

4 EGSE in the late Nineties

The development of the Rosetta mission of the European Space Agency started in the nineties. The space probe consists of two parts: the Rosetta orbiter and the Philae lander. The journey started toward ***Comet* Churyumov–*Gerasimenko*** on 2nd March, 2004.

The Philae lander is the first set of research equipment in the history of space exploration that gently descends to a comet core where it can investigate changes in the activity of a comet. All the equipment on the lander is connected to the CDMS (Command and Data Management System) which is the central data acquisition and control computer of the lander. Communication goes through the orbiter to the Earth.

During tests the system had five dedicated computers and units developed for EGSE system. Figure 6 shows the onboard bus simulator unit of EGSE for RPC (Rosetta Plasma Consortium) and its block-scheme seen on Figure 5 [4].

The RTI (Rosetta Telemetry Interface) simulator is an embedded processor system with its own embedded software and its own embedded processor and data clock line (131 kHz) common for command/telemetry, plus an on-board clock line, and several other signals. This is the interface toward the Rosetta Onboard Bus. The communication toward the EGSE PC goes on another, RS-232 bus. This interface is galvanic isolated by opto-couplers because RTI simulates the Rosetta Orbiter onboard bus with both its command bus and telemetry bus. The received telemetry data are converted and stored in RAM that are sent to EGSE PC upon software request. Commands, prepared by EGSE software are sent to RTI, stored in RAM, and converted to serial packets according to Onboard bus standard, and sent to it serially.

The graphical user interface was developed in LabWindows CVI of National Instrument development environment and it runs in Windows XP. A similar EGSE hardware configuration was designed for CONSERT instrument however, the graphical user interface software was different.

During the implementation we had to meet many requirements. The attached Figure 7 shows the grounding solution between EGSE and of the measured device. The galvanic isolation protects the electronics from any errors spreading.

The CDMS (Command and Data Management System) is the onboard computer of Rosetta Lander. The name of Rosetta Lander is Philae. It was designed by our institute in cooperation with SGF Ltd.

The Rosetta Lander Simulator carries the check of CDMS.

Tasks of Rosetta Lander Simulator are the following:

- staff training,
- testing operational schedules,
- performing long term tests,
- performing endurance tests,
- performing data transfer tests,
- running and testing telecommand sequences,
- testing software of the onboard computers, and
- reproduction of events recorded from the probe.

Complex tasks of testing are distributed among five computers in the lander simulator are as follows: one computer is used by operator to steer simulation and archive results for evaluation. Another one simulates the onboard data handling. It is connected to CDMS of the RPS bus simulator through RS-232 line. The other three computers simulate the following interfaces:

1. PC:
    a. Power Sub System (PSS)

   b. Thermal Control Unit (TCU),

2. PC:
   a. Active Descent System (ADS)
   b. Landing Gear (LG)
   c. Anchor
   d. Sampling and Drilling System (SD2),
3. PC:
   a. Scientific equipment (APX, CIVA/ROLIS, CONSERT,COSAC, MUPUS, PTOLEMY, ROMAP, SESAME),
4. PC controls the telecommand and telemetry simulator
5. PC controls the all simulation and archives measured data.

The application of 5 PCs make possible to follow all service data parallel on different displays at the same time.

During system design, the main aspect was flexibility. Besides current application this system can be adapted to simulate other complex systems. The modular structure of the system provides the possibility for developers to work on modules simultaneously and mainly independently from each other. During a long development phase involving international cooperation, e.g. the project Rosetta, its design flexibility has shown to be an important advantage. The XML based script language makes it easy to define simulations without changing the source code of any of the programs. The creation of the simulation script files does not require advanced programming skills from the different operators who are involved in the project during the long term of the mission. The software elements of the special tasks are mainly independent from the simulated system. In the case of development of difficult modules there is an opportunity to use a C++ API which supports the developers to integrate a new complex module easily into the system. Rosetta Lander Simulator is a shared computer network consisting of five computers. CDMS message handler is based on a transputer. The control PC software was developed in C language (National Instrument LabWindows). It controls the PWC activity using an XML file to ensure user friendly environment for operators.

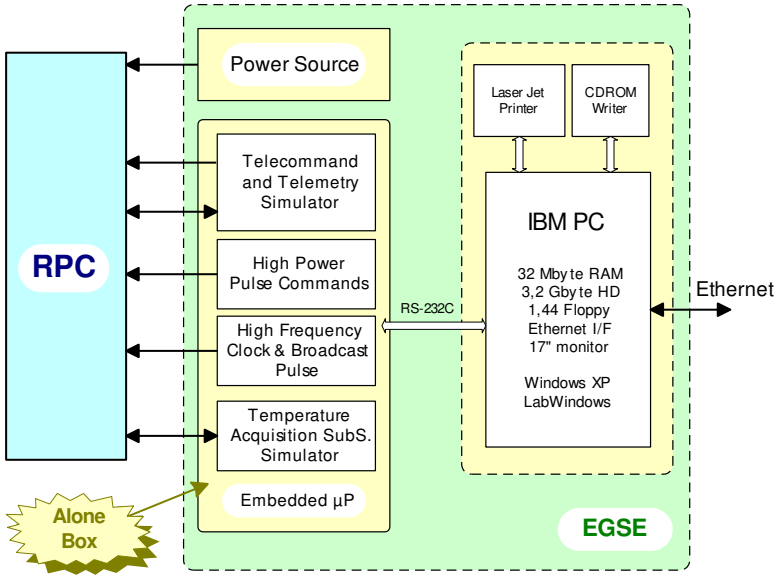


Figure 5
The Blockscheme of TEGSE for the RPC instrument for Rosetta

Figure 6
Photo of the realized EGSE

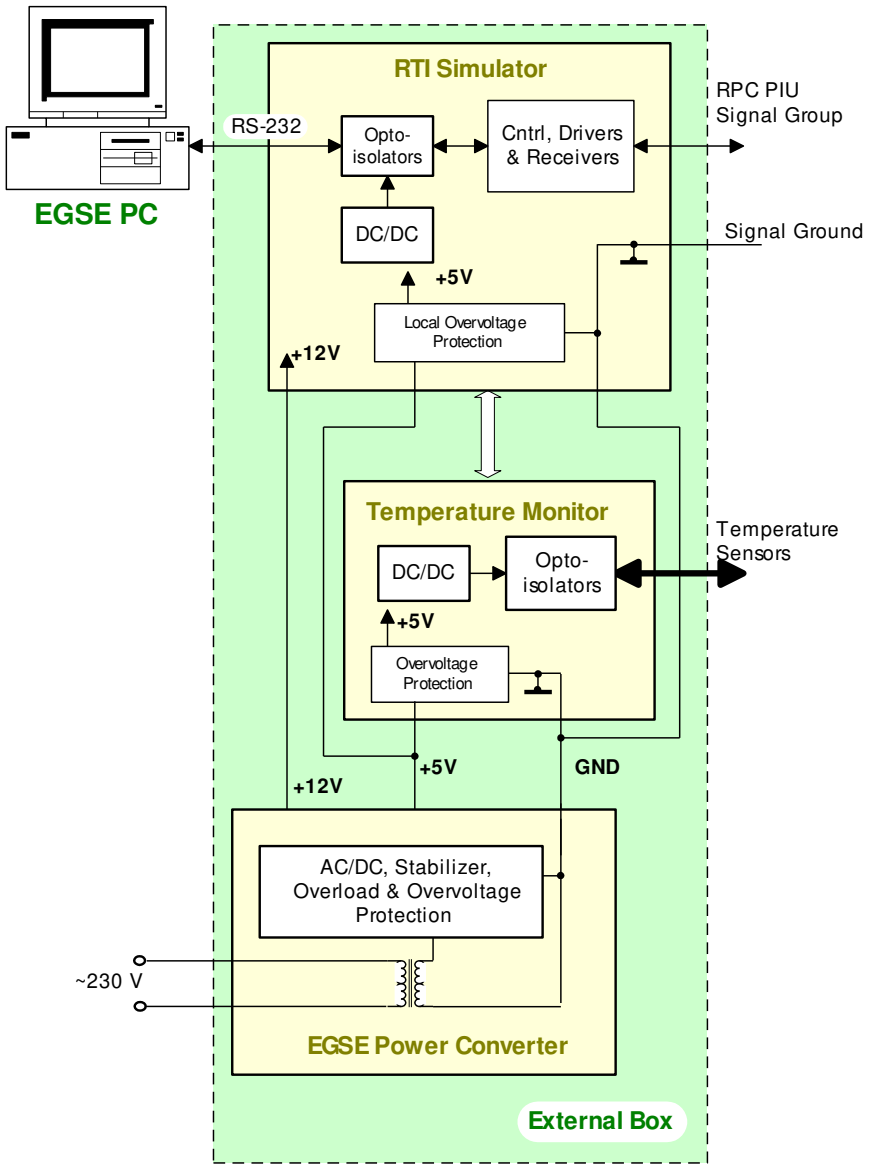


Figure 7
Grounding solution of EGSE for Rosetta orbiter

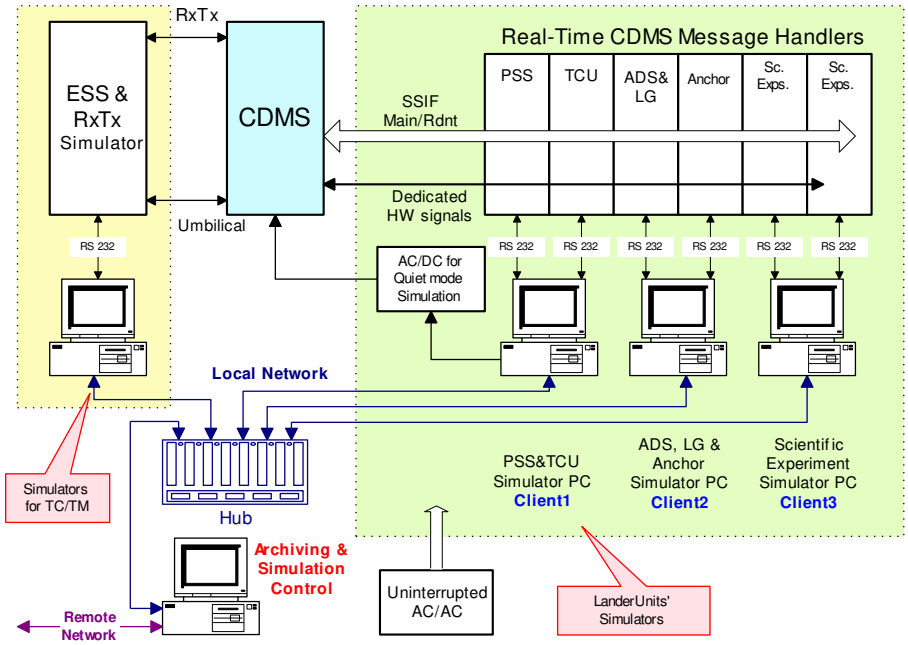


Figure 8
Blockschme of test environment of CDMS

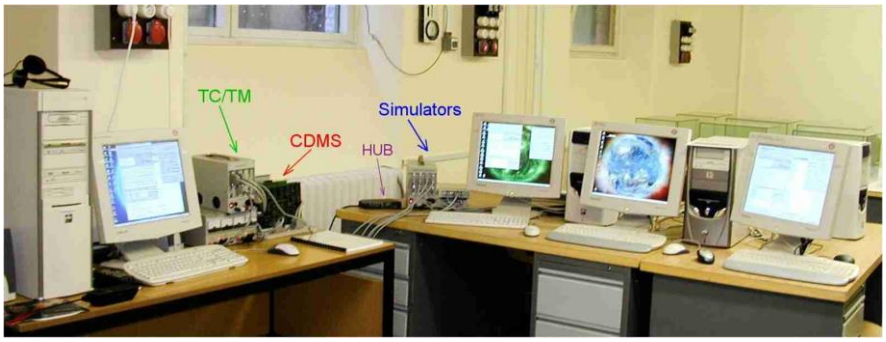


Figure 9
Lab appearance during testing of the fault tolerant central computer for Philae

5 EGSE after 2000

Our institute and SGF Ltd. participate in the development of the Obstanovka, (Obstanovka is a Russian word, it means enviroment, the other term for it is PWC (Plasma Wave Complex) was also used) system to measure particle and electric environment of the ISS (International Space Station). The PWC contains three computers working as a distributed intelligence system and eleven sensors. One

computer BSTM (Block of Storage of Telemetry Information Unit) is located inside the ISS (International Space Station), two other ones called DACU1 (Data Acquisition and Control Unit) and DACU2 with connected sensors are outside on two branches [5, 6, 7]

The widespread application of embedded processors has enabled engineers to integrate processors on individual cards. It improves intelligence and computational power of any particular card or unit.

The full checkout of Obstanovka requires several functional units, power supply units and communication channel simulators (onboard Ethernet network, amateur radio channel, bit serial data acquisition system and the so-called analog monitoring system). The EGSE has to simulate the data flow of sensors, too. Simulators have to represent real hardware interfaces. Generally the EGSE of any onboard data acquisition system has four interfaces:

1. User interface to monitor and control the system (display and keyboard);
2. Instrument (space craft) interface, realized on dedicated hardware elements;
3. Data flow source (data simulators of sensors, most cases dummy data flow is satisfactory);
4. Network interface to distribute and archiving the TM (Telemetry) data flow (Ethernet).

The PWC-EGSE system simulates the data traffic of the experiments and ISS onboard equipment connected to the PWC computers BSTM, DACU1 and DACU2. The EGSE system consists of an embedded PC104 computer producing the data traffic in real-time, and a connected User InterFace computer (UIF). This commercially available computer displays data sent to the ISS onboard system, enables switching of power supplies and sends commands and parameters to the experiments upon user interaction. The EGSE for Obstanovka (and for its data acquisition and control computers) consists of two main units: a commercially available computer PC with Ethernet interface, and a stand-alone box which contains an ISS signals simulator part (OMTC Onboard Monitoring Telemetry Interface signals) and simulators of sensor units. The standalone box realizes a low level simulation of signals connecting to the BSTM and DACUs units. This low-level signal simulator box contains a removable hard disk drive (HDD), enabling offline telemetry data read-out and provides for the possibility of preparing measuring control sequences to be delivered onboard. The PC implemented software code enables the EGSE to process and analyze housekeeping and science data either in real-time or from archives in off line mode. The delivered configuration has adequate storage capability for temporary data storage, while permanent data storage is performed by the UIF computer. The possible sensor stimulators are not part of the EGSE, they are provided by the experimenter teams.

The Onboard Monitoring Telemetry Interface (OMTC) unit has four different types of data acquisition channels:

1. "Analogue housekeeping" data monitoring system simulator;
2. Bit serial digital interface (special serial bus);
3. Amateur radio interface channel;
4. ISS Ethernet network.

Data stream acquired by the instrument interface unit is transferred to PC via Ethernet communication channel. The sensor simulators send out adequate signals for BSTM and DACUs. The OMTC simulator and the sensor simulators are built in a common box. The functional units of this stand-alone box are shown in the Figure above.

An embedded processor controls both simulators. The processor unit is built on an Intel type microprocessor running a real-time multitasking Linux based operating system. The embedded processor and the UIF PC are connected through Ethernet using TCP/IP protocol. The standalone box can be used also as the instrument interface of the Obstanovka system, excluding the sensor simulator part.

The "user interface" program runs on the PC under Windows operating system. It is a graphical interface to control the system activity and to visualize the telemetry data flow. The software was developed in C language using the National Instrument's LabWindows/CVI development tool.

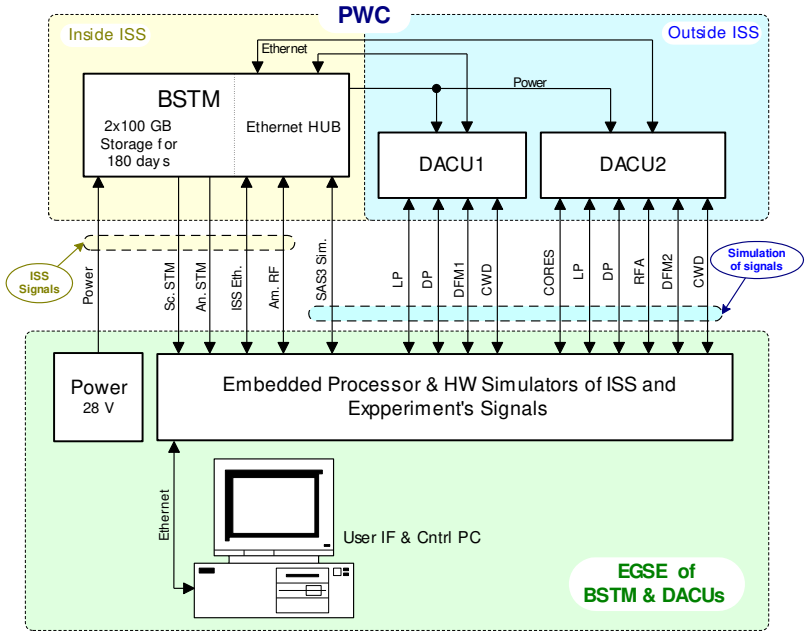


Figure 10
Test arrangement of Obstanovka computers with EGSE during test

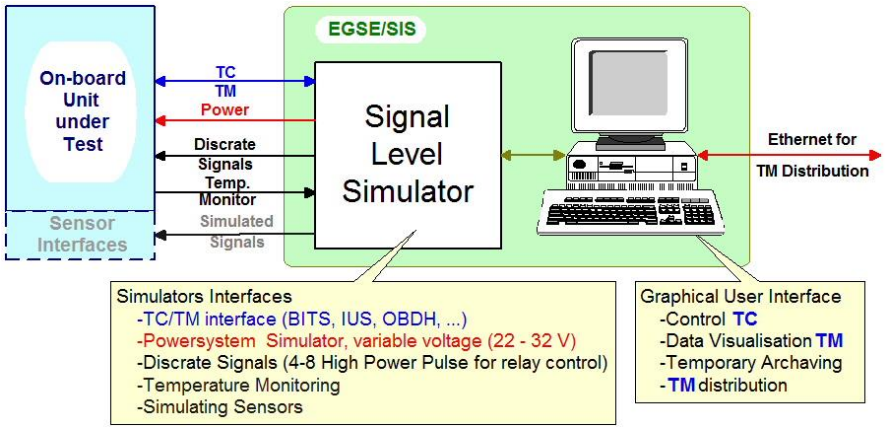


Figure 11
Detailed functional Blockscheme of EGSE

Figure 12
Photo of realized EGSE of Obstanovka

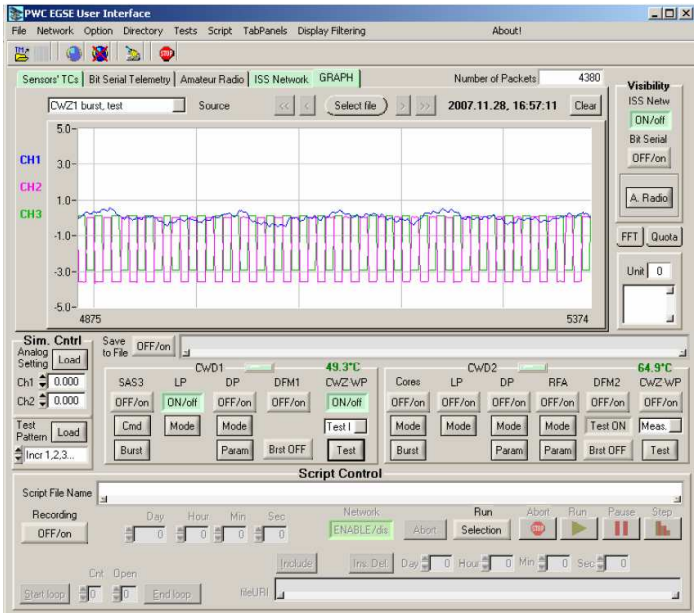


Figure 13
Screen of EGSE software, the picture shows the test of CWZ WP sensor

Application EGSE for OBSTANOVKA and Results

Two sets of EGSE for Obstanovka were used in Budapest and two other sets in Moscow. Their task included testing devices, proving functionality and finding accidental malfunctions.

The main functions of EGSE makes a wide range examination of flight hardware possible, which we present shortly. The User Interface of EGSE is based on the so-called panel (like windows) oriented graphical interface (Figure 13).

By using different areas of the GUI panel the operator can make a selection of control instructions in order to perform a desired action. Operator can perform multi-sided investigation by EGSE software features which are the following:

- save, open and decode TM (telemetry) and HK (housekeeping) data files and set IP adddresses of Ethernet channels,

- can select data received through different communication channels, (Sensor TM, BITS (Bit Serial System), Amateur Radio, onboard Ethernet) for vizualisation on display.

- visibility control can be activated or deactivated to simulate connection state between onboard channels and Earth receiver station,

- if sensors are simulated by EGSE different simulated data patterns can be selected in order to be sent by the simulated sensor,- sensor control buttons to power on/off real sensors or simulated sensors,

- Script Control, a series of commands to set working modes of sensors can be written in a file and commands can be executed by the given timing,

- quota is in connection with TM data archiving, determines the memory size for sensor data. Qouta can be switched off for test purposes,

- FFT button displays the spectrum window and enables the Fast Fourier Transformation for CWD1 (Combined Wave Sensor), CWD2, and DFM1 (Flux gate magnetometer) devices,

- TM flow archiving (save to file),

- preloaded binary commands, each of these files describes a typical command sequence of BSTM. In case of a typical command sequence a preloaded command file can be executed from the GUI with one TC instead of sending the commands of the typical sequence one by one,

- EGSE can perform data distribution through TCP/IP server port. User program running on EGSE can not only save data in files and display them according to the filtering item setting, but EGSE program can run like a server and forward measured data towards network from where they can be achieved by other computers connected to Ethernet. The other computer can receive data of any experiment by setting the IP and port address.

During test procedure after inspection several physical parameters, power transients at switching, EGSE test starts with investigating data transmission through the five communication channels of onboard devices. There are three Ethernet channels and two others using special protocol, they are BITS and OMTS (Onboard Monitoring Telemetry System). The channels are driven by data streams selected by the operator who investigates data traffic on PC screen and the waveforms with scope, too.

The EGSE can simulate high precision onboard clock and operator examines whether the onboard computers can synchronize their inner clocks to onboard clock by reading message of BSTM. Synchronizing the OBSTANOVKA clock to exact onboard clock is important for reconstruction place of data acquisition in orbit around the Earth, based on orbit parameters and saved sampling time.

EGSE software makes possible to test the onboard computers simulating sensors and sensors can be examined by it, too.

The same tests are repeated by placing the OBSTANOVKA components in thermo-vacuum chamber where the pressure is below 10^{-4} milibar. The mechanical fastening surface temperature is regulated as if it were the contact surface of ISS. The vacuum test lasts for a week since temperature transients are slow.

After vibration stress the electronic tests are repeated, too.

As usual in space developments different models of OBSTANOVKA were built, they are Technological Model, Engineering Model (identical with Flight Model) and Flight Model. All of them went through a long test procedure with EGSE here, in Budapest and the equipment was tested with EGSE in Moscow for months and after successful EGSE tests with maker of the Russian Segment of ISS.

After the successful verification procedure OBSTANOVKA was carried onto ISS by a Progress spaceship on the 11th February, 2013. The devices were placed in order through a six-hour spacewalk on 19th April, 2013. After the placement OBSTANOVKA worked properly. However, a failure occurred six months after placement onto the external wall of ISS. The operating temperature of one of sensors began to warm up slowly for two weeks and later it stopped working. Later the power supply unit of DACU1 had a sudden drop out for an unknown reason, and then it started to work again, afterwards, the analogue-digital converter fell out with its corresponding sensors, but LP1, and DP1 continued to work.

Evaluation of measurement data provided by the OBSTANOVKA is still going on, during its operation many interesting phenomena were revealed.

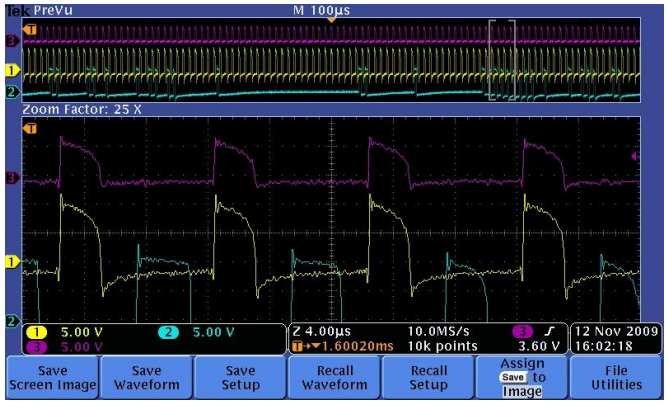


Figure 14
Examination of BITS signals

Figure 15
Assembly of OBSTANOVKA on external wall of ISS during six-hour spacewalk

Conclusions

EGSE (Electrical Ground Support Equipment) is the test device developed in order to support the development and test of flight units. EGSE task in space engineering is to simulate onboard interfaces connecting with space instruments.

The tasks they perform are:

- simulation of onboard Power Supply,
- simulation of Interfaces,
- functional test of units,
- possibility to measure important parameters,
- availability of “go” or “not go” test,
- to ensure verification within possible range of operation,
- receiving and evalutation data.

As, new computer technologies and hardware elements are applied in research instrument development the naturally have to be applied in EGSE development, too. The first EGSE devices were standalone dedicated systems of which development required a lot of engineering capacity. Involvement of PC, industrial standard buses with ready made industrial cards and later application of embedded computers enhanced effectiveness and the realisation of sophisticated intelligent systems.

The advancement of hardware was followed by that of software. In stand alone systems of the eighties the operation system was individually mostly developed in assembly language. Application of industrial standards accelerated the software development by applying Windows, Linux and C language in preparation of EGSE software.

Application of embedded systems converted EGSE in distributed intelligence systems in which the operator can control all functional tests of research equipment and signal level simulation of sensors. Adavantages of EGSE after 2000 are presented in the OBSTANOVKA experiment. OBSTANOVKA EGSE contains PC and embedded processor for control and evaluation of data and for real-time simulation onboard environment and sensors to ensure comprehensive and true to real application test.

Acknowledgement

Our jobs were supported by the Hungarian Space Office and the NKFI A (National Research, Development and Innovation Found).

References

[1] R. Z. Sagdeev, F. Szabó, G. A. Avanesov, P. Cruvellier, L. Szabó, K. Szegő, A. Abergel, A. Balázs, I. V. Barinov, J. L. Bertoux, J. Blamont, M. Detaille, E. Demarelis, G. N. Dulnev, G. Endrőczy, M. Gárdos, M. Kanyó, V. I. Kostenko, V. A. Krasikov, T. Nguyen-Trong, Z. Nyitrai, I. Rényi, P. Rusznyák, V. A. Shamis, B. Smith, K. G. Sukhanov, S. Szalai, V. I. Tarnapolsky, I. Tóth, G. Tsukanova, B. I. Valnicek, L. Várhalmi, Yu. K. Zaiko, S. I. Zatsepin, Ya. L. Ziman, M. Zsenei, B. S. Zhukov: Television Observation of Comet Halley from VEGA Spacecraft, Nature Vol. 32l, l5 May l986, pp. 262-266

[2] S. Szalai: The Imaging System on Board the VEGA Spacecraft, Images of the Nucleus of Comet Halley, ESA SP-1127, 1996, Vol. 2, pp. 20-32

[3] Balázs A., Breuer P.,Erényi I., Gárdos M., Hamza E., Kovács G., Pongrácz J., Rényi I., Szalai S.: A VEGA TV-rendszer földi támogatása. Mérés és Automatika, 33. évf., l985. l-2 szám, 38-52. old.

[4] S. Szalai, A. Balazs, A. Baksa, G. Tróznai: Rosetta Lander Software Simulator, 57th International Astronautical Congress, Valencia, Spain, 2006 (on DVD of 57 IAC)

[5] Horváth, Lipusz, Nagy: Űrkutatás – magyar részvétel a Nemzetközi Űrállomáson, adatgyűjtő és vezérlő számítógép az Obsztanovka-kísérlethez, Elektronet 2004/4 91-93

[6] Balajthy K., Szalai S.: A nemzetközi űrállomásra kerülő „Obsztanovka" kísérlet földi ellenőrző berendezése, Elektronet, 2004 8. szám, 19-20 oldal

[7] Adatgyűjtő és vezérlő számítógép a Nemzetközi Űrállomás Obsztanovka kísérletéhez Balajthy Kálmán, Endrôczi Gábor, Nagy János *KFKI Részecske- és Magfizikai Kutatóintézet,* Horváth István, Lipusz Csaba, Dr. Szalai Sándor *SGF Kft.,* Híradástechnika 2006/4 17-22

Platform for Computer-aided Harmonization of Informatics Curricula

Milinko Mandić[1], Zora Konjović[2], Mirjana Ivanović[3]

[1] Faculty of Education, University of Novi Sad, Podgorička 4, 25000 Sombor, Serbia, milinko.mandic@pef.uns.ac.rs

[2] University of Singidunum, 32 Danijelova St., 11 000 Belgrade, Serbia, zkonjovic@singidunum.ac.rs

[3] Faculty of Science, University of Novi Sad, Trg DositejaObradovića 3, 21000 Novi Sad, Serbia, mira@dmi.uns.ac.rs

Abstract: This paper presents a new platform aimed at improving informatics teaching by computer-aided harmonization of the standardized secondary school informatics curriculum and curricula by which teachers of informatics are educated. The platform relies on competency based curricula ontologies and the harmonization method based on ontology alignment. The secondary school informatics curriculum ontology was built to comply with the ACM K12 standard, while the teachers' curriculum ontology was built based on selected existing curricula, due to the lack of explicit standardization in the field. A task-specific method for curricula harmonization was developed that relies on standard ontology alignment algorithms. The prototype software tool was implemented and used by independent experts to verify the proposed method, by investigating compliance of the standardized secondary school informatics curriculum and the domain (informatics) segment of the teachers' curriculum.

Keywords: Informatics education; curriculum; ontology alignment; ACM K12

1 Introduction

The research presented in this paper was motivated by well recognized needs for frequent and even substantial changes in informatics teaching curricula at primary and secondary education levels caused by the extreme dynamics of changes in the informatics field and its complexity, along with labor market increasing IT competences requirements regarding all professions and all qualification levels. This gives an important role to existing IT competences and shifts the educational paradigm "from an input-centered approach to an output-focused student-centered approach" [1]. In order to keep pace, curricula for educating informatics teachers must be changed to respond by ensuring the necessary teachers' competences.

Hence, the representation of informatics curricula for educating informatics teachers and informatics curricula of lower levels of education is needed as well as tools that will facilitate curricula changes while keeping them compliant in terms of required informatics teachers' competences.

The rest of the paper is organized as follows. The second section presents related work. Section three presents briefly, the proposed ontological models of the curriculum for educating informatics teachers and the informatics curriculum for secondary education level. Section four presents the procedure underlying the software tool for curricula harmonization. The fifth section presents verification of the proposed platform by means of investigation of the compliance of the standardized secondary school level informatics curriculum and the domain (informatics) segment of the proposed teachers' curriculum. Finally, the sixth section contains concluding remarks, which include an evaluation of the achieved results and directions for further research.

2 Related Work

In accordance with the research presented in this paper (informatics curricula harmonization by ontology matching with an emphasis on acquired competences), the papers dealing with the application of ontology for the representation of the curricula and papers dealing with ontology matching and its applications to curricula harmonization were analyzed.

Ontological approaches are increasingly being applied to represent curricula, since ontology is machine-readable, reusable and sharable [2] [3] [4] [5]. Ontologies can represent the educational domain from different perspectives [6] [7], providing "a richer description and retrieval of learning contents" [2]. According to [3], ontologies are most appropriate for the development of curricula based on intended learning outcomes, students' competence and standards. In [4], a proposal for an ontology curriculum in the field of computing is provided and an idea of applying ontologies is described by which the user can choose from a drop down menu the desired learning outcome and, in accordance with the selected outcome, the corresponding concepts in the ontology developed are labeled. In [2], ontologies are applied as a basis of software for the development and maintenance of an educational curriculum that provides information on the length of instructional units, the duration of instruction, assessment instruments and the display of untaught lessons and the like. Demartini et al. [5] present an ontology representing the academic environment as suggested by the Bologna reform. The proposed ontology does not contain an explicit representation of the curriculum. Gluga et al. [8] describe a system that models curriculum design in university teaching programs. The system exploits a lightweight semantic mapping approach to map learning goals from multiple accrediting sources across the degree. In [9], a system for representing ACM CS curriculum based on the IEEE RCD standard is shown.

A range of different techniques and strategies for ontology alignment have been implemented in a number of systems, as is evident in [10] [11] [12]. Despite wide use of ontologies' application for representing curricula, as well as numerous publications dealing with researches of matching and alignment of ontologies such as [13] [14], in contemporary literature one can rarely find examples of implemented systems for the alignment of ontological representations of the curricula (different or the same levels of education). In [15] the authors emphasize the importance of a system for harmonizing curricula that have been modeled using ontologies. Conceptual maps were created describing the curricula translated into an ontology, where algorithms for alignment of study programs were neither described nor implemented.

3 The Ontological Model of Curricula

The main goal of the research presented in this paper was to propose a tool that would help in determining whether teacher education curriculum provides the competencies required for teaching in a high school. Therefore, the models of teacher education and secondary school informatics curricula are based on competencies and as such, the base class of both ontological models is *Competence*. Numerous definitions of competence [16] [17] [18] all agree with what is presented in [19], i.e., that the notion of competence, regardless the context, refers to successfully performing a task or activity, that is adequate acquaintance of some domain's ***knowledge*** or ***skill***. Therefore, in this paper, the knowledge and skills mapped to specific classes of an ontological model curriculum (*Knowledge* and *Skills*), are represented as subclasses of *Competence* as described in detail in [20]. Thematic areas of the curriculum are mapped to subclasses of the *Knowledge* class, whereas the skills acquired through the study of specific subject areas are mapped to the corresponding subclasses of the *Skills* class. The *Skills* subclasses and the *Knowledge* subclasses are related via the object property *hasKnowledge*, that is its inverse property *hasSkill*. To ensure interoperability with learning management systems that provide information about competence, upper ontology classes are modeled in accordance with the IEEE RCD standard as described in [9].

Analysis of the content and form of teacher education curricula available on the web sites of institutions in several countries (Germany, Austria, Turkey and the Republic of Serbia) shows that competencies corresponding to each subject (course) are determined primarily by two fields: *course content* and *course outcome*. In our model of curriculum *course content* corresponds to the *Knowledge* class and *course outcome* to the *Skills* class. Skills are represented by classes corresponding to the categories of the cognitive process dimension of the revised Bloom's taxonomy [21], which is the dominant taxonomy in the area of CS and in general [22]. Exceptions are 'remember' and 'understand' categories,

which are represented by a single class *Remember-understand*. Thus, the *Skills* subclasses are: *Remember-understand*, *Apply*, *Analyze*, *Evaluate* and *Create.*

Since no proposals of standardized curricula models for informatics teachers' education exists yet, an ontological model of a teacher education curriculum was created based on our analysis of 22 teacher education curricula from different countries (Germany, Austria, Israel, Estonia, Turkey, Scotland, USA and R. of Serbia), as well as the recommendations suggested by [23] [24]. Five general areas that all curricula for informatics teacher preparation must include are: Informatics (domain) knowledge, General pedagogical knowledge (educational psychology, didactics, etc.), Knowledge of the methods of teaching informatics, Knowledge of teaching practice, General knowledge (foreign languages, mathematics, the application of ICT in the realization of teaching). These five general areas were modeled by subclasses of the class *Knowledge*.

The hierarchical structure of the upper subclasses of the *Informatics_domain_knowledge* class is based on the classifications shown in [4] [25] [26]. The ontological model includes all areas of informatics knowledge contained in most of the analyzed curricula. In the ontological model of the teacher education curriculum descriptions of classes were further mapped to labels. Subclasses of the *Skills* class were created primarily based on ISTE standards specified in [24] [27]. *Skills* subclasses were also based on the outcomes/objectives of the courses contained in the analyzed teacher education curricula. Based on [21] [28], all the described teaching skills were classified in the appropriate subclasses of Bloom's taxonomy classes and then associated with the knowledge to which they can relate.

The ontological model of secondary school informatics curriculum in this paper was designed strictly following competences designed for the secondary level of education (K8 or higher levels of standard) of the ACM K12 CS curriculum proposal [29]. The ontological model of the secondary school informatics curricula is created in two phase as described in detail in [20].

Using the tool Protégé (http://protege.stanford.edu/), OWL ontologies representing high school and teachers' informatics curricula are created, which are available at addresses www.pef.uns.ac.rs/SecondaryInformaticsCurriculum/index.html and www.pef.uns.ac.rs/InformaticsTeacherEducationCurriculum/index.html respectively.

4 Method for Curricula Harmonization

For two ontologies O_1 and O_2, matching implies the process of finding an appropriate entity from O_2 for each entity from O_1. Alignment of ontologies is the output of the matching process and comprises a set of "correspondences" [13] between ontologies.

Since the object and data type properties are predefined in advance and are the same in both ontologies modeling curricula, the proposed method for curricula harmonization compares only classes of ontologies, so the harmonization model can be formally written as follows.

If the ontologies modeling two curricula are $O_1$and O_2 , C_{ik} is an ontology class, (=), (⊇), (⊆) are equivalence, one-to-many superset/superclass and one-to-many subset/subclass relations respectively and conf_i is degree of confidence, then the curricula harmonization model is

$$\text{Alignment}(O_1, O_2) = \left\{ \left(C_{i1}, C_{j2}, \text{conf}_i, \text{relation}_i \right) \middle| \, C_{i1} \in O_1, C_{j2} \in O_2, \text{conf}_i \in [0,1], \text{relation}_i \in \{=, \subseteq, \supseteq\} \right\}.$$

Figure 1 shows the diagram of the method proposed in this paper for matching the secondary school and teacher education curricula.

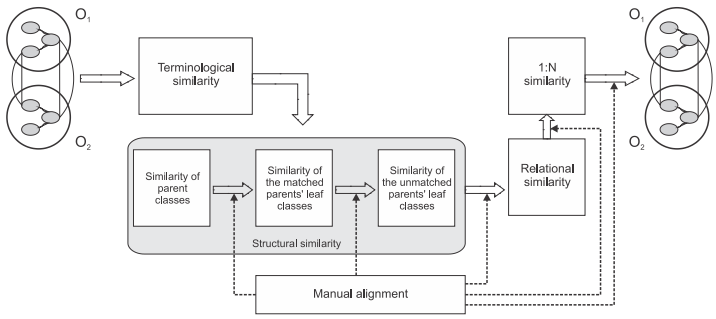


Figure 1
The procedure of matching secondary school and teacher education curricula

The matching is done in two phases. The first phase, which can be considered as pre-processing, determines terminological similarity by means of linguistic and string-based method [13] applied to local names and the classes' labels. The obtained similarity matrix is input to the second phase, which consists of the sequential composition of matchers determining structural, relational and one-to-many similarities respectively. Each matcher of this phase provides input (similarity matrix) to the subsequent matcher. The best matched pairs of classes are determined by applying the greedy selection algorithm as described in [30]. Three matchers that calculate structural similarity compare only subclasses of the *Knowledge* class in teachers' curriculum to which topics from domain (informatics) knowledge are mapped with subclasses of *Knowledge* class of secondary school curriculum because classes that belong to non-informatics knowledge in teacher education curriculum appear in teachers' curriculum only. Matching of skills structures (subclasses of the *Skills* class) is determined through relational similarity with an aim to check whether the secondary school skills are at the lower or the same level of Bloom's taxonomy with matched teaching skills.

User manual intervention is enabled after each matching stage, except after the terminological stage. Following manual interventions are enabled that preserve the consistency of one-to-many relations cardinality (e.g. superset/superclass and subset/subclass) produced by the matcher:

a) Changes in the greedy algorithm's threshold values
b) Disconnection of the matched classes
c) Changing the correspondence degree of matched classes
d) Replacement of the class in the matched pair
e) Creating a new matched pair of classes

The rest of this section contains descriptions of applied alignment algorithms and rationales for the choice of algorithms.

4.1 Terminological Similarity

Terminological similarity is determined by applying standard linguistic method based on the WordNet lexical database to strings that identify particular class. Labels are used for the additional description of concepts in the curricula; thus, when comparing classes of two ontologies using a terminological matcher, local class names and their labels are taken into account.

The similarity between two tokens belonging to the local names of classes is determined using the Lin information-theoretic similarities [31] in instances where there are two tokens in the WordNet dictionary. If this is not the case, token similarity is determined using the Jaro-Winkler method [32] [33]. Applying the greedy selection method to a matrix consisting of the similarities of all possible pairs of tokens of compared names of classes, a list S_{ln} is obtained that contains similarities of the best matched pairs of tokens. The total similarity of local names for the two classes $s_{ln}(C_{i1},\ C_{j2})$ is calculated as:

$$s_{ln}(C_{i1}, C_{j2}) = \frac{2 \cdot \sum_{i=o}^{m} S_{ln}(i)}{|tok_{i1}| + |tok_{j2}|};\ |tok_{ik}| - \text{ \# of tokens in local name of } C_{ik};\ m - \text{dimension of } S_{ln}$$

The similarity of classes' labels $s_{lb}(C_{i1},\ C_{j2})$ and the similarities between the local name of the class of one ontology and the label of the class of other ontology ($s_{lnlb}(C_{i1},\ C_{j2})$ and $s_{lbln}(C_{i1},\ C_{j2})$) are calculated analogously. The total terminological similarity for classes $s_{term}(C_{i1}, C_{j2})$ is:

$$s_{term}(C_{i1}, C_{j2}) = \max(s_{ln}(C_{i1}, C_{j2}), s_{lb}(C_{i1}, C_{j2}), s_{lnlb}(C_{i1}, C_{j2}), s_{lbln}(C_{i1}, C_{j2}))$$

4.2 Structural (taxonomic) Similarity

Structural (taxonomic) similarity is calculated in three steps:

- Calculating the similarities of all parent classes

- Calculating the similarities of non-parent (leaf) classes that are subclasses of matched parent classes
- Calculating the similarities of non-parent classes that are subclasses of unmatched parent classes

Such composition of structural algorithms enables manual intervention in order to support early correction which is necessary because the results of subsequent matchers depend on the results of the previous ones.

4.2.1 Determining the Similarity of Parent Classes

For the similarity of parent classes a slight modification of the algorithm presented in [11] is used.

For two parental classes C_{i1} and C_{j2}, the similarities of their superclasses ("parents"), the similarities of their subclasses ("children") and their terminological similarity are taken into account. There are observed similarities of *all parents and children*, not only of direct ones. Similarity between the subclasses of C_{i1} and C_{j2}, denoted by $s^{sub}(C_{i1}, C_{j2})$, is determined by the following algorithm:

/* Let A_{ij} be a class of an ontology, $A_{i1} \in O_1$ and $A_{i2} \in O_2$

If $\nexists A_{i1} | A_{i1} \subseteq C_{i1}$ or $\nexists A_{i2} | A_{i2} \subseteq C_{j2}$ then

$s^{sub}(C_{i1}, C_{j2}) = 0$

else

Let $\{A_{k1}\} \subseteq C_{i1}, k = 1, n;\ n \geq 1$ and $\{A_{l2}\} \subseteq C_{j2}, l = 1, m;\ m \geq 1$

for k = 1 to n

for l = 1 to m

/* $s_{term}(A_{k1}, A_{l2})$ are the values of similarity of classes from the set $\{A_{11} \dots A_{n1}\}$ with classes from the set $\{A_{12} \dots A_{m2}\}$

submatrix[k][l] = $s_{term}(A_{k1}, A_{l2})$

/* the list of best-matched pairs of subclasses S^{sub} is obtained applying the greedy selection method to the submatrix

S^{sub} = Greedy_Selection_Method (submatrix)

/* $s^{sub}(C_{i1}, C_{j2})$ is set to the average value of similarities of *matched* subclasses

$s^{sub}(C_{i1}, C_{j2}) = \frac{\sum_{l=o}^{p} S^{sub}(l)}{p}$, p= size of S^{sub}

The similarity of superclasses $s^{sup}(C_{i1}, C_{j2})$ is calculated analogously using similarities of each superclass of the C_{i1} class with each superclass of the C_{j2}class, and calculating the average of the matched superclasses. Overall similarity

$s_{parent}(C_{i1}, C_{j2})$, is calculated as the average of the terminological similarities and previously calculated similarities of superclasses and subclasses, provided that both classes C_{i1} and C_{j2} have at least one subclass; if the condition is not met overall similarity is 0. Modification of algorithm [11] takes place if one of compared classes has no parent class. In this case, value $s^{sup}(C_{i1},\ C_{j2})$ is omitted when calculating average for $s_{parent}(C_{i1}, C_{j2})$. That way the impact of the structural similarity is relaxed, leaving larger number of potentially useful classes for further matching which is reasonable taking into account the fact that teachers' and high school curricula have relatively different structures. The similarity matrix of this structural matcher S_{parent} has mxn dimension where *m* and *n* are the total number of *Knowledge* subclasses in ontologies O_1 and O_2, respectively. The list of matched classes A_{parent} is obtained by applying the greedy selection algorithm to the matrix S_{parent}. The similarities of predefined classes (*Knowledge, Competence*) are not taken into account in these calculations.

4.2.2 Determining the Similarities of the Matched Parents' Leaf Classes

At this stage, the similarity $s_{leaf}(C_{i1}, C_{j2})$is calculated as follows:

/* Let A_{ij} be the class of the ontology, $A_{i1} \in O_1$and $A_{j2} \in O_2$.

/* Further, let the following apply: C_{i1} is a leaf class of ontology $O_1$and C_{j2} is a leaf class of ontology O_2, or C_{i1} is a leaf class of ontology O_1 and the C_{j2} class has only leaf subclasses, or C_{j2} is a leaf class of ontology O_2 and C_{i1} has only leaf subclasses.

If $\exists\{A_{l1}, A_{k2}\} | \{A_{l1}, A_{k2}\} \in A_{parent}, A_{l1} \in \{A_{11} \dots A_{n1}\}, C_{i1} \subseteq \{A_{11} \dots A_{n1}\}, A_{k2} \in \{A_{12} \dots A_{m2}\}, C_{j2} \subseteq \{A_{12} \dots A_{m2}\}$then

$$s_{leaf}(C_{i1}, C_{j2}) = s_{term}(C_{i1}, C_{j2})$$

else

$$s_{leaf}(C_{i1}, C_{j2}) = s_{parent}(C_{i1}, C_{j2})$$

In order to avoid elimination of potentially equivalent classes that are not described with the same level of detail (by subclasses), in addition to the comparison of leaf classes, the comparison of non-leaf classes having only leaf subclasses with the leaf classes is also done.

4.2.3 Determining Similarities of the Unmatched Parents' Leaf Classes

The similarity of leaf classes C_{i1} and C_{j2} becomes zero, if no matching of the parents of C_{i1}, with any parent of C_{j2} is obtained by applying the first two structural matchers. This, together with curriculum description, which is far from being unambiguous for non-standardized curricula, could leave some essentially related concepts (with different parents), unpaired. For example, in the secondary school curriculum model the concepts of computer graphics are represented as subclasses of the *Multimedia* class, while in many teaching curricula, concepts

relating to computer graphics and those relating to multimedia, belong to distinct courses, i.e., in the teacher education curriculum, computer graphics concepts are represented as special subclasses of the *Graphics* class, while there is a separate parent class *Multimedia* containing no computer graphics concepts at all. Since the class *Multimedia* of the secondary school curriculum model is not matched with the *Graphics* class of the teacher education curriculum model, but with the *Multimedia* class, the previous matcher would calculate zero similarity measure between classes to which, for example, concepts of raster images are mapped. This problem is resolved here by explicitly defining disjointed parent classes, i.e., the classes *Multimedia* and *Graphics* of the teachers' curriculum are not defined as disjoint. Then, the principle for determining the similarity of classes $s_{disj}(C_{i1}, C_{j2})$ is as follows:

/* Let A_{leaf} be a list of matched classes obtained by a matcher that determines the similarity of leaf classes of matched parents.

/* Let the following apply:$\{\{A_{11}, A_{12}\} \dots \{A_{n1}, A_{n2}\}\} \in A_{leaf}$,$C_{i1} \subseteq \{A_{11} \dots A_{n1}\}$, $C_{j2} \subseteq \{B_{12} \dots B_{m2}\}$, $A_{k2} \in \{A_{12} \dots A_{n2}\}$,$B_{k2} \in \{B_{12} \dots B_{m2}\}$

If C_{i1} and C_{j2} are unmatched leaf classes and $\nexists A_{k2}$, B_{k2} defined as disjoint classes and $\nexists\{A_{l1}, B_{k2}\} \mid \{A_{l1}, B_{k2}\} \in A_{leaf}$,$A_{l1} \in \{A_{11} \dots A_{n1}\}$, $B_{k2} \in \{B_{12} \dots B_{m2}\}$ then

$$s_{disj}(C_{i1}, C_{j2}) = s_{term}(C_{i1}, C_{j2})$$

else

$$s_{disj}(C_{i1}, C_{j2}) = s_{leaf}(C_{i1}, C_{j2})$$

This matcher in the sequential composition is after the matcher determines the similarity of matched parents' leaf classes, which favors matched parents' classes, but also extends the search space to other non-disjoint classes that could contain some useful concepts.

All structural matchers calculate similarities only between the *Knowledge* subclasses, so the similarities of the subclasses of the *Skills* class are not changed by structural alignment step.

4.3 Determining Relational Similarity

The outcomes/objectives of the course or subject areas in our ontological models are simply mapped to the corresponding subclasses of Bloom's taxonomy classes, which are the subclasses of the *Skill* class. This makes determination of the similarity of classes on the basis of their taxonomic structure inappropriate for this part of the ontology. On the other hand, the outcomes of the curricula (mapped to the appropriate *Skills* subclasses in the ontology) are usually described by a larger free text, which makes the use of only a terminological matcher inappropriate. Therefore, the similarity of *Skills* subclasses in the system is calculated based on the relation graph. The method for calculating relational similarity applied in the

paper is based on the principle used in [34]: *if the two classes that represent the domains of object properties (relation) are similar, and if the object properties are also similar, then the classes representing the ranges of the domain classes are similar* [13]. Relational similarity $s_{rel}(C_{i1}, C_{j2})$ is determined as follows:

/*Let A_{disj} be a list of matched classes obtained by a matcher that determines the similarity of leaf classes of unmatched parents

/* Let $C_{Knowledgei}$ be the *Knowledge* class and let $C_{Skillsi}$ be the *Skills* class

If $C_{i1} \subseteq C_{Knowledge1}$ or $C_{j2} \subseteq C_{Knowledge2}$ then

$$s_{rel}(C_{i1}, C_{j2}) = s_{disj}(C_{i1}, C_{j2})$$

else if $C_{i1} \subseteq C_{Skills1}$ and $C_{j2} \subseteq C_{Skills2}$ then

If C_{i1} is associated with $\{A_{11} \dots A_{n1}\} | \{A_{11} \dots A_{n1}\} \subseteq C_{Knowledge1}$ and C_{j2} is associated with $\{A_{12} \dots A_{m2}\} | \{A_{12} \dots A_{m2}\} \subseteq C_{Knowledge2}$ then

If $\{A_{k1} \dots A_{l1}\}$ is the set of all superclasses and subclasses of all classes from $\{A_{11} \dots A_{n1}\}$, $k = n + 1$ and $\{A_{o2} \dots A_{p2}\}$ is the set of all superclasses and subclasses of all classes from $\{A_{12} \dots A_{m2}\}$, $o = m + 1$ then

If $\exists\{A_{q1}, A_{r2}\} | \{A_{q1}, A_{r2}\} \in \mathrm{A_{disj}}$, $A_{q1} \in \{A_{11} \dots A_{n1}\} \cup \{A_{k1} \dots A_{l1}\}$, $A_{r2} \in \{A_{12} \dots A_{m2}\} \cup \{A_{o2} \dots A_{p2}\}$ then

$$s_{rel}(C_{i1}, C_{j2}) = s_{term}(C_{i1}, C_{j2})$$

else

$$s_{rel}(C_{i1}, C_{j2}) = 0$$

If a structure exists in the part of the ontology to which the subclasses of the *Skills* class belong (some outcomes are further structured), then for these subclasses, when calculating a relational similarity, the relations inherited from their superclasses are taken into consideration. Due to the fact that in our model, the object property that connects *Knowledge* and *Skills* subclasses is known and the same in both ontologies, "the circularity" which could be caused by using the relational method [13] is reduced (the similarity of object properties based on the similarity of the domain and range is not explicitly calculated).

4.4 Determining 1:N Similarity

Previously described algorithms determine to what extent the classes of ontology O_1 are *equivalent* to the classes of ontology O_2 with cardinality of 1:1. The next alignment phase enables matching of a class of one ontology with *multiple classes* of the other ontology through relation *superclass/subclass*. The following pseudo-code describes the method that determines whether some class C_{i1} from O_1 is a superclass of classes from O_2.

/* Let A_{rel} be a list of matched classes obtained by a matcher determining the relational similarity

If $\{C_{i1}, C_{j2}\} \in A_{rel}$ and $\nexists A_{l1} | A_{l1} \subseteq C_{i1}$ and $\exists A_{k2} | A_{k2} \subseteq C_{j2}$ then

If $\nexists \{A_{l1}, A_{k2}\} | \{A_{l1}, A_{k2}\} \in A_{rel}, A_{l1} \in O_1,\ A_{k2} \in \{A_{12} \dots A_{n2}\},\ \{A_{12} \dots A_{n2}\} \subseteq C_{j2},\ n \geq 1$ then

$$\{A_{12} \dots A_{n2}\} \subseteq C_{i1}$$

An analogous procedure is applied to determine whether the class C_{j2} is a superclass of classes from O_1. After applying this method, a class can be associated with several classes of the other ontology by superclass and equivalence relations. Conversely, a class can be a subclass of the ontology class to which it belongs, as well as, the class of the other ontology.

5 Verification of the Proposed Curricula Harmonization Method

Based on the models and algorithms described in Sections 3 and 4 of this paper, the software application for curricula harmonization was implemented using the Java programming language. Evaluation of the software was carried out by the expert team composed of 4 university professors in the field of informatics teacher education, 2 employees in the Education District Offices (Ministry of Education) and 2 teachers teaching secondary school informatics. Their tasks were to define the reference alignment and to interpret the results. In the rest of this section the results obtained by the software tool application to the curricula from Section 3 and the experts' analysis of these results are presented following the matching steps (matchers) applied after terminological matching.

5.1 Similarity of the Parent Classes

Figure 2 shows a part of the matched classes of compared curricula obtained by the first taxonomical/structural algorithm that determines the similarity of classes that have at least one subclass, with the threshold set to 70%.The percentage of matched *Knowledge* subclasses at this stage was 14.9%.

The column "Source class" and "Target class" contain the local names of classes of ontological representations of secondary school and teacher education curricula, respectively; the column "Type of relation" identifies the type of relation between the classes (Equivalence, Superclass and Subclass), while "Similarity Value" denotes the correspondence value between the matched classes.

The expert team noticed that certain classes with identical names were matched with the similarity value below 100% and that some classes were matched despite

not having similar names (Figure 2). The explanation for this is the presence of an additional description in the labels of teacher education curriculum for some classes and/or the participation of similarities of superclasses/subclasses in the calculation of the overall similarity of classes.

Row	Source class	Target class	Type of relation	Similarity...
1	Algorithms	Algorithms	Equivalence	83.18%
2	Connections_Between_Mathematics_and_Comp...	Mathematical_basis_of_informatics	Equivalence	76.35%
3	Data_Structures	Data_types_and_structures	Equivalence	91.31%
4	Databases	Database	Equivalence	89.32%
5	Fundamentals_of_Hardware_Design	Memory	Equivalence	75.33%
6	Hypertext_Language_HTML_tags	HyperText_Markup_language_-_HTML	Equivalence	86.7%
7	Levels_of_Language_Software_and_Translation	Programming_paradigms	Equivalence	70.02%
8	Models_of_Intelligent_Behavior	Artificial_intelligence	Equivalence	74.31%
9	Multimedia	Multimedia	Equivalence	85.42%
10	Object-oriented_programming	Object-oriented_programming	Equivalence	94.52%
11	Parts_of_a_Computer	Hardware_basics	Equivalence	75.07%
12	Phases_of_the_software_development_process	Models_and_phases_of_the_software_deve...	Equivalence	89.52%
13	Principles_of_Software_Engineering	Software_engineering	Equivalence	90.78%
14	Principles_of_computer_organization	Architecture_and_Organization	Equivalence	79.44%
15	Problem_Solving_and_Algorithms	Problem_solving	Equivalence	89.84%
16	Problem_solving	Problem_solving_phases	Equivalence	84.13%
17	Programming_Languages	Programming_Fundamentals	Equivalence	92.03%
18	Representing_Information_Digitally	Data_representation	Equivalence	77.02%
19	Structured_programming	Structured_and_Imperative_programming	Equivalence	87.57%
20	The_major_component_parts_of_the_microproc...	Central_processing_unit_-_CPU	Equivalence	77.29%
21	Web_Page_Design_and_Development	Web_technologies_and_development	Equivalence	76.76%

Figure 2
Matched classes after applying the algorithm for parent classes matching

In addition, it was found that some classes having the same names in the secondary school curriculum and teacher education curriculum (for example, *Problem solving*) were not mutually matched, but that the *Problem_solving* class of the secondary school curriculum and the *Problem_solving_phases* class in the teacher education curriculum were matched (row 16); the expert team considered this as correct, because the subclasses of both matched classes represent stages in algorithmic problem solving.

Additionally, looking only at the names of the matched classes from Figure 2, the matching of the classes *Levels_of_Language_Software_and_Translation* and *Programming_paradigms* (row 7) could be considered as false. However, the topics of secondary school and teacher education curricula (differences and comparison of high level languages and machine languages, levels of programming languages, etc.) described by their subclasses are corresponding.

At this level of the application of a structural matcher, the expert team identified a pair of incorrectly matched classes *{Fundamentals_of_Hardware_Design, Memory} (*row *5).* However, since their parent classes were correctly matched, this pair of classes does not influence the similarity of their subclasses, which will be calculated by the following matchers.

5.2 Similarities of the Matched Parents' Leaf Classes

Figure 3 displays some matched classes obtained after applying a taxonomic/structural algorithm that determines the similarity between leaf classes of the matched parents. The percentage of matched *Knowledge* subclasses at this stage was 61.18%.

The similarity of the matched classes obtained at this stage was determined by the terminological similarity of their local names and labels, under condition that some of their parent classes were matched by the matcher calculating the similarity of parent classes, which explains why classes *Repetition* and *Iteration* were highly matched (Figure 3, row 17). Namely, their non-direct parent classes *Programming_Languages* and *Programming_Fundamentals* had already been matched (Figure 2, row 17). Further, since the verbs *repeat* and *iterate* are considered as synonymous within the WordNet database, the terminological matcher showed high similarity for the *Repetition* and *Iteration* classes.

An example of matching a leaf class to a class that is the parent of leaf classes is the match {*Knowledge-based_Systems, Semantic_Web_and_knowledge_representation*} (Figure 3, row 12). The class *Knowledge-based_Systems* has no subclasses and is a subclass of the *Models_of_Intelligent_Behavior* class. The class *Semantic_Web_and_knowledge_representation* has subclasses and is a subclass of the *Artificial_intelligence* class matched with the class *Models_of_Intelligent_Behavior* by applying the matcher for calculating the similarities of parent classes (Figure 2, row 8).

R...	Source class	Target class	Type ...	Similar..
1	Careers_related_to_computers	Profession_and_Careers_in_computing	Equiv...	83.15%
2	Challenges_of_modeling_information_digitally	Representation_of_the_different_types_of_informati...	Equiv...	71.78%
3	Client_side_scripts_in_a_networked_environment	Client-side_scripting	Equiv...	75.0%
4	Code_a_solution_from_a_design	Software_deployment	Equiv...	80.0%
5	Conversion_among_decimal_binary_and_hex_number_sy...	Conversion_among_different_number_systems	Equiv...	93.33%
6	Creating_a_web_site_that_conforms_to_standards	Basic_Principles_of_creating_web_sites	Equiv...	77.46%
7	Diagnose_and_troubleshoot_PC_problems	Maintenance_and_support_of_PC_hardware	Equiv...	70.02%
8	Encoded_data_and_integrated_circuits	Characteristics_of_digital_integrated_circuits	Equiv...	77.44%
9	Hardware_to_support_multimedia	Hardware_supporting_multimedia	Equiv...	100.0%
10	Interactivity	Static_and_Dynamic_web_content	Equiv...	76.89%
11	Interface_evaluation	Measures_for_evaluation_in_HCI	Equiv...	77.44%
12	Knowledge-based_Systems	Semantic_Web_and_knowledge_representation	Equiv...	80.16%
13	Natural_Language	Natural_language_processing	Equiv...	80.0%
14	Presentation_software	Software_applications_for_presentations	Equiv...	80.0%
15	Relationships_among_high-level_languages_assembly_l...	Higher_level_languages_vs_machine_level_langua...	Equiv...	73.35%
16	Relevancy_of_web_sources	Sharing_documents_on_the_web	Equiv...	71.0%
17	Repetition	Iteration	Equiv...	100.0%
18	Routing_protocols_for_connection-communication	Principles_of_routing	Equiv...	85.93%
19	Using_the_clipboard	Functions_of_interrupts	Equiv...	72.22%
20	What_is_Intelligence	Definition_of_artificial_intelligence	Equiv...	85.71%

Figure 3

Example of matched classes after applying the second structural algorithm

At this stage, the expert team reported substantially incorrect matches (row 16, 19), which were true candidates for manual interventions.

5.3 Similarities of the Unmatched Parents' Leaf Classes

According to the previous matcher, some subclasses of the *Multimedia* class *(Create_edit_and_save_bitmapped_images, Vector_versus_bit-mapped_images, Create_edit_and_save_vector_images)* of the secondary school curriculum had not been matched with subclasses of the *Multimedia* class of the teacher education curriculum. By applying the algorithm for calculating the similarities of the leaf classes of unmatched parents, these classes were matched with the subclasses of the *Graphics* class (Figure 4).The percentage of matched *Knowledge* subclasses at this stage was 82.35%.

...	Source class	Target class	Type of rela...	Simila...
1	Create_edit_and_save_bitmapped_images	Programs_to_create_and_edit_raster_graphics	Equivalence	79.5%
2	Create_edit_and_save_vector_images	Programs_to_create_and_edit_vector_graphics	Equivalence	81.49%
3	Vector_versus_bit-mapped_images	Methods_of_presenting_static_images_in_com...	Equivalence	100.0%

Figure 4

Matched leaf classes whose parents were not paired

Classes that remain unmatched after the application of the structural algorithm may indicate incompleteness of knowledge in the teacher education curriculum or incompatible structures of curricula ontologies. Examples of incompleteness in teacher education curriculum correctly detected by the system are machine cycle phases, robotics, documentation techniques and elements of user friendly software. Example of false incompleteness detected in the teacher curriculum, which is caused by incompatible structures of the curricula ontologies, are those related to connections between mathematics and computer science where the unmatched class *Functions_including_parameters_and_mathematical_notation* in the secondary school curriculum is a subclass of the class *Connections_between_mathematics_and_computer_science*, while in the teacher education curriculum corresponding knowledge was mapped to a subclass of the *General_knowledge* class that does not belong to the CS domain knowledge at all. Finally, differences in the structure of ontologies arising from the depth of studying specific topics in the secondary school and teacher education curricula may result in unmatched classes that do not necessarily point to an inadequate teacher education curriculum. An example is the thematic area of the secondary school curriculum 'Interdisciplinary Utility of Computers and Problem Solving in the Modern World' with focuses representing the various applications of computers including 'Education and Training'. Since these focuses were mapped to the leaf subclasses of the class *Interdisciplinary_utility_of_computers_and_problem_solving_in_the_modern_world* in the secondary school curriculum, despite the fact that the teacher education curriculum contains classes (such as *Educational_software* and *E-learning*) that correspond to the focus 'Education and training*'* from the secondary school curriculum, these classes were not matched with the leaf class *Education_and_training*, due to the fact that in the teacher education curriculum they have class structures not considered by the proposed matchers.

5.4 Relational Similarity

In terms of the lowly-structured subclasses of the *Skills* class (practically the only structure by which Bloom's taxonomy is modeled), where the titles and labels of subclasses usually contain free text, terminological matching significantly affects the final results. To avoid omitting potentially useful matches that can be used for manual intervention, in this instance, a lower criterion (threshold) was set in the determination of the matched classes (60%). Percentage of paired classes was 80.88%.

A part of the results obtained using the relational matcher determining the similarity between the subclasses of the *Skills* class is shown in Figure 5. The "Bloom" column in the table in Figure 5contains the ⊤ mark if the level of skill in the teacher education curriculum is higher or equal to the level required in the secondary school curriculum, or the ⊥ mark if not.

The opinion of the expert team was that some matched classes here are potentially inaccurate (rows 3, 7, 11 and 14). The classes that were not matched because there was no corresponding class in the teacher education curriculum were the classes *Explain_the_relationship_between_a_web_server_a_web_page_and_a_browser* and *Describe_the_difference_in_the_processing_of_arrays_stacks_and_queues*.

R...	Source class	Target class	Ty...	Simil...	Bloom
1	Convert_a_word_problem_into_code_using_top-down_design	Design_programs_in_languages_from_two_different_programming_par...	Eq...	71.45%	⊤
2	Convert_between_decimal_binary_and_hexadecimal_numbers	Apply_arithmetic_operations_in_different_number_systems	Eq...	70.04%	⊤
3	Convert_between_image_formats	Contrast_vector_and_raster_graphics	Eq...	77.24%	⊤
4	Create_a_Web_site_given_design_specifications	Design_web_pages	Eq...	69.16%	⊤
5	Create_a_user-centered_design	Design_interactive_user_interfaces_for_diverse_applications	Eq...	74.59%	⊤
6	Define_intellectual_property_and_state_the_impact_of_provisions_to_prote...	Discuss_intellectual_property	Eq...	64.16%	⊤
7	Define_parallel_processing	Use_design_patterns	Eq...	61.94%	⊤
8	Describe_the_major_applications_of_artificial_intelligence_and_robotics.	Apply_Artificial_intelligence_applications	Eq...	70.7%	⊤
9	Describe_the_role_of_the_OS_as_an_intermediary_between_application_...	Explain_the_objectives_and_functions_of_modern_operating_systems	Eq...	64.59%	⊤
10	Design_a_multi-table_relational_database	Project_relational_data_model	Eq...	75.0%	⊤
11	Determine_if_a_given_algorithm_successfully_solves_a_stated_problem	Select_basic_language_instructions_to_accomplish_a_given_straightfor...	Eq...	64.34%	⊥
12	Display_a_multimedia_object_within_a_Web_page_or_document	Set_the_multimedia_on_the_web	Eq...	61.21%	⊤
13	Evaluate_algorithms_by_their_efficiency_correctness_and_clarity	Analyze_algorithms_using_complexity_efficiency_aesthetics_and_correct...	Eq...	69.93%	⊥
14	Evaluate_computer_components_in_terms_of_features_and_price	Understand_machine_level_components_and_related_issues_of_comp...	Eq...	65.63%	⊥
15	Express_the_design_of_a_Web_site_using_standard_tools	Use_web_design_tools	Eq...	72.73%	⊤
16	List_ways_to_increase_computer_performance	Propose_options_to_improve_computer_performance	Eq...	67.85%	⊤
17	Name_and_explain_the_steps_in_the_problem-solving_process	List_problem_solving_phases	Eq...	72.81%	⊤
18	Name_the_different_phases_of_the_software_development_process	Use_one_or_more_software_development_models	Eq...	69.61%	⊤
19	Use_modeling_and_simulation_to_represent_and_understand_natural_ph...	Use_Modeling_and_simulation_to_solve_real_world_problems	Eq...	74.53%	⊤
20	Utilize_advanced_OS_user_interface_elements_and_features	Use_interactive_graphic_OS	Eq...	62.91%	⊤
21	Write_conditional_statements_that_include_simple_and_complex_Boolean...	Create_complex_logical_expressions_using_Boolean_operators_and_f...	Eq...	75.93%	⊤

Figure 5

A part of matched skills of the secondary school and teacher education curricula

The expert also reported that some outcomes in the secondary school curriculum were represented by a larger number of skills subclasses than the corresponding outcomes in the teacher education curriculum. Consequently, some skills from the secondary school curriculum remain unpaired, even when the teacher education curriculum contains classes that include these skills (such as *Code_a_program_to_solve_a_stated_problem_using_variables_and_at_least_one_decision_or_loop* and *Use_advanced_search_engine_options_and_refine_searches_to_locate_information*).

5.5 1: N Similarity

An example that justifies application of the 1:N algorithm is the matching of the subclasses of the *Semantic_Web_and_knowledge_representation* class and the *Knowledge-based_Systems* class. Since the class *Semantic_Web_and_knowledge_representation* contained unmatched leaf subclasses and the *Knowledge-based_Systems* leaf class was matched with *Semantic_Web_and_knowledge_representation* (Fig 3, row 12), the system suggested the 1:N relation, i.e., that the subclasses of the *Semantic_web_and_knowledge_representation* class (*Ontology, Predicate_logic, Web_ontology_language,* etc.) could also be the subclasses of the *Knowledge-*

based_Systems class (Figure 6).The total percentage of matched *Knowledge* subclasses achieved after the last matching phase was 87%.

...	Source class	Target class	Type of rela...	Similari...
1	Knowledge-based_Systems	Knowledge_representation_in_educa...	Superclass	80.16%
2	Knowledge-based_Systems	Ontology	Superclass	80.16%
3	Knowledge-based_Systems	Predicate_logic	Superclass	80.16%
4	Knowledge-based_Systems	Proposotional_logic	Superclass	80.16%
5	Knowledge-based_Systems	Resource_Description_Framework_-...	Superclass	80.16%
6	Knowledge-based_Systems	Semantic_web_-_basic_notions	Superclass	80.16%
7	Knowledge-based_Systems	Web_ontology_language	Superclass	80.16%

Figure 6
Matched classes in "Superclass" relation

5.6 Prototype Performance and Usability

Performance measures *Precision* (0.64), *Recall* (0.76) and *F-measure* (0.695) were obtained using reference alignment derived by human experts and results obtained by matching system, which is in accordance with reference [35] that gives maximum importance to the recall measure when ontology alignment is a semi-automatic process.

The expert team evaluated these results as acceptable. They also found the tool useful "as it is" for improving concrete teacher education curriculum in order to meet the requirements of the ACM K12 curriculum. The acquired class pairs evaluated as incorrect justify the need for the semi-automatic method for curricula harmonization.

The obtained quantitative results about the percentage of matched classes and the preliminary evaluation imply that the model of the teacher education curriculum is satisfactorily harmonized with the ACM K12 model. Still, the experts reported that even preliminary results obtained by means of the software prototype correctly indicate some subject areas that are not covered by the model of teacher education curriculum (machine cycle phases, documentation techniques, robotics, user-friendly web design, Interface evaluation, etc.) and that the teacher education curriculum does not provide all the skills needed for teaching in accordance with the ACM K12 curriculum proposal. Therefore, it is necessary to improve the teacher education curriculum so that it represents the missing knowledge and skills. In addition, some of the unmatched classes indicate incompatible structures of the ontological models. Typical examples are 'Connections Between Mathematics and Computer science' and 'Interdisciplinary Utility of Computers'. Such information makes a system useful for improvement of structure of the teacher education curriculum model. Also, some *Skills* classes of the ACM K12 model remained unmatched even in the teacher education model: there is the *Skills* class that could be considered as their superclass. Consequently, it is necessary to improve the teacher education curriculum so that the skills related to programming and the use of Internet be described in more detail/with a greater number of classes.

Conclusions and Future Work

The focus of this paper is the task-specific semi-automated method which can assist in development and maintenance of the teacher education curricula as to provide teachers' competences required by changes in the high school informatics curricula.

OWL ontologies of standardized secondary school informatics curriculum and the curriculum for the education of informatics teachers were developed, where the ontology of the secondary school curriculum relies on the ACM K12 standard, while the ontology of the teacher education curriculum was designed on the basis of representative informatics teachers' education curricula. The ontological models for both curricula have the same top level of competencies model (classes *Knowledge* and *Skills*) and the same relational structure (*hasKnowledge, hasSkill*). The task-specific semi-automated method based on standard algorithms for ontology alignment for curricula comparison was proposed, and a software tool prototype was developed supporting the proposed method. Using the software prototype and curricula ontologies, the team of experts consisting of university professors in the field of informatics teacher education, employees of the Education District Offices (Ministry of Education) and teachers teaching secondary school informatics carried out verification of the proposed approach by means of investigation of the compliance of the standardized secondary school curriculum with the teacher education curriculum.

There are two advantages of the proposed curricula model. The first one is machine readable representation of both curricula that facilitates exchange and joint development of curricula, while the second one is its capacity to support representation of the standardized curricula, which is confirmed by ontology representing ACM K12 compliant secondary school curriculum. The constraints are model's capacity to represent some important additional curriculum aspects (instructional design, teaching materials, etc.) and its heavy reliance upon competences not being easy to define unambiguously. The latest is confirmed by experts reporting that the values of similarity, as well as the adequacy of matching, were lower in classes modeling the outcomes/skills of subject areas or courses. Extending ontologies as to comprise other curriculum aspects could alleviate the first constraint, while the second one could be alleviated by better structuring the ontology part that represents skills and/or by utilizing fuzzy ontologies. Future research concerning curriculum model will take these directions.

The main advantages of the proposed curricula harmonization method are the utilization of the standard ontology alignment methods for curricula comparison modified as to exploit the model of competences common to both curricula, and manual intervention option available to experts that could provide for acquiring and integrating deeper experts' knowledge into curriculum model. The need for manual intervention option is already confirmed by independent experts' reports

indicating that some of the class pairs obtained at certain stages do not reflect the real similarity between equivalent concepts in the curricula. The constraints are close coupling of the method with the ontological model and performance issues. The architecture of the matching engine enables simple introduction of other types of matchers (like internal structural similarity or extensional methods) and/or modification of the existing ones in accordance with ontological model thus relaxing the first constraint. One way to improve performance is to apply some procedures for the early elimination of matching candidates. Future research regarding the system's performance will also explore the possibilities of using the approach described in [36]. Last but certainly not least important, a further research direction is the improvement of the evaluation by means of increasing the set of curricula to be evaluated and extending the experts team.

References

[1] Adam, S. (2008) Learning Outcomes, Current Developments in Europe: Update on the Issues and Applications of learning Outcomes Associated with the Bologna Process. *Bologna Seminar: Learning outcomes based higher education: the Scottish experience*. http://www.ehea.info/Uploads/Seminars/Edinburgh_Feb08_Adams.pdf

[2] Fernández-Breis, J. T., Castellanos-Nieves, D., Hernández-Franco, J., Soler-Segovia, C., Robles-Redondo, M. C., González-Martínez, R. & Prendes-Espinosa, M. P. (2012) A Semantic Platform for the Management of the Educative Curriculum, *Expert Systems with Applications*, 39(5), 6011-6019

[3] Dexter, H., & Davies, I. (2009) An Ontology-based Curriculum Knowledgebase for Managing Complexity and Change. *Ninth IEEE International Conference on Advanced Learning Technologies*, *ICALT*, 136-140

[4] Cassel, L., Davies, G., LeBlank, R., Snyder, L., & Topi, H. (2008) Using Computing Ontology as a Foundation for Curriculum Development. *Proc. SWEL@ITS '08: The Sixth International Workshop on Ontologies and Semantic Web for E-Learning*, 21-30

[5] Demartini, G., Enchev, I., Gapany, J., & Cudré-Mauroux, P. (2013) The Bologna Ontology: Fostering Open Curricula and Agile Knowledge Bases for Europe's Higher Education Landscape. *Semantic Web - Interoperability, Usability, Applicability*, 4(1), 53-63

[6] Alatrish, E. S., Tošić, D., & Milenković, N. (2014) Building Ontologies for Different Natural Languages. *Computer Science and Information Systems*, *11*(2), 623-644

[7] Vesin, B., Ivanović, M., Klašnja-Milićević, A. & Budimac, Z. (2013) Ontology-based Architecture with Recommendation Strategy in Java Tutoring System. *Computer Science and Information Systems*, *10*(1), 237-261

[8] Gluga, R., Kay, J. & Lever, T. (2013) Foundations for Modeling University Curricula in Terms of Multiple Learning Goal Sets. *IEEE Trans. on Learning Technologies*, *6*(1), 25-37

[9] Mandić, M., Segedinac, M., Savić, G., & Konjović, Z. (2013) IEEE RCD Standard-based Ontological Modeling of Computer Science Curriculum. *Proceedings of the 3rd International Conference on Information Society and Technology*, 189-285

[10] Cruz, I., Stroe, C., Caimi, F., Fabiani, A., Pesquita, C., Couto, F., & Palmonari, M. (2011) Using AgreementMaker to Align Ontologies for OAEI 2011. *Proceedings of the Sixth International Workshop on Ontology Matching*, 114-121

[11] Jean-Mary, Y. R., Shironoshita, E. P., & Kabuka, M. R. (2009) Ontology matching with semantic verification. *J. Web Semantics*, *7*(3), 235-251

[12] Li, J., Tang, J., Li, Y., & Luo, Q. (2009) RiMOM: A Dynamic Multistrategy Ontology Alignment Framework. *IEEE Transactions on Knowledge and Data Engineering*, *21*(8), 1218-1232

[13] Euzenat, J., & Shvaiko, P. (2007) *Ontology Matching*, Springer-Verlag, Berlin-Heidelberg, p. 333

[14] Shvaiko, P., & Euzenat, J. (2011) Ontology Matching: State of the Art and Future Challenges. *IEEE Trans. Knowl. Data Eng.*, *25*(1), 158-176

[15] Anohina-Naumeca, A., Graudina, V., & Grundspeņķis, J. (2012) Curricula Comparison using Concept Maps and Ontologies. *International Scientific Conference: Applied Information and Communication Technologies*, 5, Jelgava (Latvia), 177-183

[16] Fleishman, E. A., Wetrogan, L. I., Uhlman, C. E., & Marshall-Mies, J. C. (1995) Abilities In Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R. & Fleishman, E. A. (Eds.), *Development of Prototype Occupational Information Network Content Model* , Salt Lake City, UT: Utah Department of Employment Security, Vol. 1, p. 1086

[17] Learning Technology Standards Committee of the IEEE Computer Society (2008) *IEEE Standard for Learning Technology—Data Model for Reusable Competency Definitions*. http://www.doleta.gov/usworkforce/pdf/2007-ieeecomp.pdf

[18] Mirabile, R. (1997) Everything You Wanted to Know about Competency Modeling. *Training and Development*, *51*(8), 73-77

[19] Shippmann, J., Ash, R., Battista, M., Carr, L., Eyde, L., Hesketh, B., Kehoe, J., Pearlman, K., Prien, E., & Sanchez, J. (2000) The Practice of Competency Modeling. *Personnel Psychology*, *53*(3), 703-740

[20] Mandić, M., Konjović, Z., & Ivanović, M. (2015) Ontological Model of the Standardized Secondary School Curriculum in Informatics. *Proceedings of*

the 5th International Conference on Information Society and Technology, 363-367

[21] Krathwohl, D. R. (2002) A Revision of Bloom's Taxonomy: An Overview. *Theory into Practice*, *41*(4), 212-218

[22] Fuller, U., et al. (2007) Developing a Computer Science-Specific Learning Taxonomy. *ACM SIGCSE Bulletin*, *39*(4), 152-170

[23] Gal-Ezer, J., & Stephenson, C. (2010) Computer Science Teacher Preparation is Critical. *ACM Inroads*, *1*(1), 61-66

[24] International Society for Technology in Education (2011) ISTE Standards Computer Science Educators. http://www.iste.org/docs/pdfs/20-14_ISTE_Standards-CSE_PDF.pdf

[25] Association for Computing Machinery (2008) ACM Computer Science Curriculum 2008: An Interim Revision of CS 2001 CS curriculum. http://www.acm.org/education/curricula/ComputerScience2008.pdf

[26] Association for Computing Machinery (2012) The 2012 ACM Computing Classification System. http://www.acm.org/about/class/2012

[27] International Society for Technology in Education (ISTE) (2002) Educational Computing and Technology Standards for Secondary Computer Science Education Initial Endorsement Program. http://www.iste.org/docs/pdfs/ncate_iste_csed_2002.pdf?sfvrsn=2

[28] Churches, A. (2007) Bloom's Digital Taxonomy http://www.techlearning.com/techlearning/archives/2008/04/andrewchurches.pdf

[29] The CSTA Standards Task Force (2011). CSTA K–12 Computer Science Standards. http://csta.acm.org/Curriculum/sub/CurrFiles/CSTA_K-12_CSS.pdf

[30] Wu, W., Yu, C., Doan, A., & Meng, W. (2004) An Interactive Clustering-based Approach to Integrating Source Query interfaces on the Deep Web. *Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*, 95-106

[31] Lin, D. (1998) An Information-Theoretic Definition of Similarity. *Proceedings of the 15th International Conf. on Machine Learning*, 296-304

[32] Jaro, M. (1989) Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, *84*(406), 414-420

[33] Winkler, W. (1999) *The State of Record Linkage and Current Research Problems*, tech. report 99/04, Statistics of Income Division, Internal Revenue Service Publication

[34] Maedche, A., & Staab, S. (2002) Measuring Similarity between Ontologies. *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web,* 251-263

[35] Stoilos, G., Stamou, G., & Kollias, S. (2005) A String Metric for Ontology Alignment. In: Gil, Y., Motta, E., Benjamins, V. R., Musen, M. A. (eds.) *ISWC 2005. LNCS,3729*, 623-637

[36] Ehrig M., & Staab, S. (2004) QOM - Quick Ontology Mapping. *Proceedings of the 3rd International Semantic Web Conference (ISWC04)*, 683-697

Adaptive Model Predictive Controller for Web Transport Systems

N. Muthukumar[1], Seshadhri Srinivasan[2], K. Ramkumar[1], K. Kannan[3], Valentina Emilia Balas[4]

[1]Electric Vehicle Engineering and Robotics Lab (EVER), SASTRA University, 613401 Thanjavur, India; muthukumar.n@sastra.ac.in; ramkumar@eie.sastra.edu

[2] International Research Center, Kalasalingam University, Srivilliputtur, India; seshadhri@ieee.org

[3]Department of Humanities and Sciences, SASTRA University, 613401 Thanjavur, India; kkannan@maths.sastra.edu

[4]Department of Automatics and Applied Informatics, Aurel Vlaicu University of Arad, B-dul Revolutiei 77, 310130 Arad, Romania; balas@drbalas.ro

Abstract: In web transport systems (WTS), parameter variations in transported material affects the product quality and integrity of the processed material. This investigation presents an adaptive model predictive controller for WTS in process industries considering the variations of the web radius with respect to time. The proposed controller uses a radius approximation algorithm for estimating the changes in web radius. The controller then updates the WTS model with the estimated changes in web radius that enters as the disturbance in the system. Then, the controller uses the model with disturbance estimation algorithm and an optimization routine to compute the future control moves that guarantee product quality, while simultaneously satisfying the physical and operating constraints of WTS. Furthermore, it assures product quality with web radius variations. The main advantage of the proposed controller is, it combines the predictive and optimality features of MPC with adaptation provided by an adaptive controller. Simulations on WTS used in paper industries illustrates the performance and advantages derived by employing the adaptive MPC against conventional MPC. Our results show a reduction in the peak overshoot, integral absolute error, and integral time square error by using the adaptive MPC.

Keywords: Process industries; Adaptive model predictive controller (AMPC); Web Transport Systems (WTS); Web Transport Controllers (WTC); parametric variations

1 Introduction

Web Transport Systems (WTS) are widely used in process industries, to transport finished material in the form of sheets over long distances. To have good material integrity and for reducing downtime due to loss of production resulting from web breaks, constant tension should be maintained on the transported material and Web Tension Controllers (WTCs) are used to this extent. Therefore, WTC performance is pivotal for having good product quality and increasing production. Parameter variations of the web material affect the WTC performance adversely. Therefore, to assure product quality, the controller should maintain the constant tension in spite of the parameter variations of the web material.

The problem of web tension control has received considerable attention and many control approaches have been proposed. Proportional integral and derivative (PID) controllers are the most widely used ones (see, [6] [7] [8] [13] [14] [15] and [18]). Although PID controllers are simple and easy to tune, their performance deteriorates due to disturbances acting on the web and parameter variations. In [2] an optimal controller was studied to optimize the product quality. To compensate the effect of parameter variations on product quality, the investigations in ([1] and [11]) proposed the use of adaptive controllers. Though, the adaptive controllers designed in these investigations performed better than optimal controllers in the presence of parameter variations, they lacked optimal performance.

More recently, the investigations in ([10] and [17]) proposed the use of model predictive controllers (MPC) for WTS. The MPC uses the model of the web transport system, estimate of disturbance and an optimization routine to optimize the product quality and energy consumption in the WTS. The results showed that, combining prediction with optimization provided significant cost and quality benefits. Though, MPC showed optimal and predictive capabilities, their performance was limited by the parameter variations in the web material. Our objective in this investigation is to overcome the shortcomings with the existing approaches by designing a controller that combines adaptive, optimal and predictive features.

To reach the objectives, this investigation proposes an adaptive MPC (AMPC) that combines the benefits of adaptive and model predictive controller. The main building blocks of the AMPC are: an online parameter estimator that computes the variations in radius (web parameter), a disturbance estimator, a constrained optimization routine that computes the future optimal outputs and a receding horizon strategy. The output of the parameter estimation algorithm updates the model, during each time epoch. The AMPC uses the updated model, an optimization routine, and knowledge of constraints to compute the control moves that improves the product quality. The optimization is performed using a receding horizon approach. The use of estimation model within AMPC

guarantees that the optimization is performed considering the underlying changes in web transport dynamics and parameter variations obtained using parameter estimation technique. The presence of disturbance and parameter estimates improves the model accuracy used for designing the MPC, and handles the physical constraints and operating constraints inherently in the design. As a result, the controller is more robust to parameter variations and other disturbances acting on the plant. The AMPC combines the adaptive, optimal and predictive control features in the existing approaches and therefore, is a promising approach for web transport systems that are subjected to frequent parameter variations.

The main contributions of the investigation are: design of adaptive MPC controller for the web transport systems that incorporates parameter estimation, a simple parameter estimation algorithm, and an illustration of the proposed control methodology using parameters obtained from a prototype web transport system. The investigation highlights the advantages in employing AMPC by comparing the obtained results with a conventional MPC.

This paper is organized into six sections. Section 2 describes the mathematical model of the WTS. Section 3 provides AMPC design. Section 4 explains the AMPC algorithm for web tension control. The results and discussion are provided in Section 5. The conclusion section is presented in Section 6.

2 Mathematical Model of WTS

This section describes the WTS and the dynamical equations of the system. Figure 1 shows a prototypical WTS employed widely in a process industry and it consists of a un-winder, winder, dancer, and gear transmission system. The dancer is connected to a potentiometer for measuring the displacement and a spring and damping arrangement is used for regulating its displacement. The parameters and variables that describe the dynamics of WTS are shown in Table 1. The winder and un-winder rollers are coupled to an electric motor through a geared drive. The motor torque is the input to WTS and it influences the roller velocity to maintain constant tension acting on the web. The load cell measures the web tension and provides feedback to WTC. The web tension is regulated by adjusting the torque of roller motors and dancer position.

The mathematical model of WTS describes the relation between the torque and the web tension. The dynamics can be described using the block diagram shown in Figure 2. The mathematical model of WTS consists of three sections, namely, *(i)* drive train, *(ii)* web material and *(iii)* dancer.

2.1 Drive Train

The relation between the motor torque and the angular velocity of the motor, and the relation between the angular velocity and the roller tangential velocity are given by (1) and (2), respectively by [4] as

$$\dot{\omega}_m = -\frac{B_m}{J_m}\omega_m + \frac{1}{J_m}\tau - \frac{1}{J_m}\tau_{cou} \tag{1}$$

$$V_2 = \frac{R}{N}\omega_m \tag{2}$$

Table 1
List of parameters and variables used in WTS model

SYMBOL	DESCRIPTION
τ	Input torque of the winder motor in Nm
J_m	Moment of Inertia of the motor in Nm^2
B_m	Viscous friction constant of the motor in Nm.s/rad
ω_m	Angular Velocity of the motor in rads/sec
R	Radius of the winder in m
N	Gear ratio
E	Young's modulus in Pa
A	Cross sectional area of the web in m^2
L	Length of the web in m
T	Tension of the web in N
V_1	Tangential Velocity of the un-winder roller in m/sec
V_2	Tangential Velocity of the winder roller in m/sec
B_d	Viscous friction of the dancer in N.s/m
M_d	Mass of the dancer in Kg
K_d	Spring constant of damper system in N/m
d	Position of the dancer in m
V_d	Velocity of the dancer in m/sec
τ_{cou}	Coupling torque in Nm
F_x	Disturbance forces acting on dancer in N

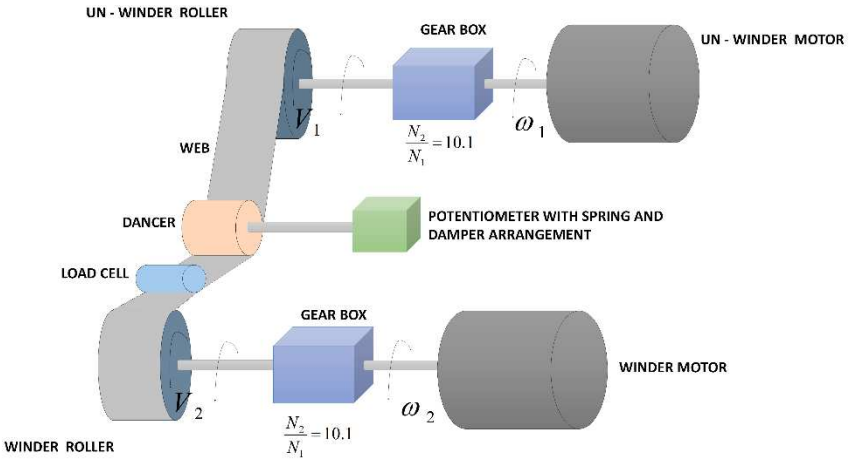


Figure 1
Schematic of WTS with un-winder, winder and dancer arrangement

2.2 Web Material

Using Hook's law and law of conservation of mass the relation between web tension (T), angular velocity (ω_m) of roller and dancer velocity (V_d) can be modelled as

$$\dot{T} = -\frac{V_1}{L}T + \frac{REA}{NL}\omega_m - \frac{EA}{L}V_1 - \frac{2EA}{L}V_d \tag{3}$$

2.3 Dancer

From Newton's second law motion, the relationship between the web tension (T) and dancer position (d) is given by

$$\ddot{V}_d = \frac{2}{M_d}T - \frac{1}{M_d}F_x - \frac{B_d}{M_d}V_d - \frac{K_d}{M_d}d \tag{4}$$

The state space model of WTS can be framed form equations (1), (3) and (4) as:

$$\overbrace{\begin{bmatrix}\dot{V}_d\\ \dot{d}\\ \dot{\omega}_m\\ \dot{T}\end{bmatrix}}^{\dot{x}} = \overbrace{\begin{bmatrix}-\frac{B_d}{M_d} & -\frac{K_d}{M_d} & 0 & \frac{2}{M_d}\\ 1 & 0 & 0 & 0\\ 0 & 0 & -\frac{B_m}{J_m} & 0\\ -\frac{2EA}{L} & 0 & \frac{REA}{L} & -\frac{V_1}{L}\end{bmatrix}}^{A}\overbrace{\begin{bmatrix}V_d\\ d\\ \omega_m\\ T\end{bmatrix}}^{x} + \overbrace{\begin{bmatrix}0\\ 0\\ \frac{1}{J_m}\\ 0\end{bmatrix}}^{B}\tau + \overbrace{\begin{bmatrix}0 & -\frac{1}{M_d} & 0\\ 0 & 0 & 0\\ 0 & 0 & -\frac{1}{J_m}\\ -\frac{EA}{L} & 0 & 0\end{bmatrix}}^{F}\overbrace{\begin{bmatrix}V_1\\ F_x\\ \tau_{cou}\end{bmatrix}}^{d_x} \tag{5}$$

$$y = \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} V_d \\ d \\ \omega_m \\ T \end{bmatrix} \tag{6}$$

where, A is the system matrix; B is the input matrix; F is the disturbance matrix; x contains the states; d_x is the disturbances; τ is the input toque applied to the WTS and $y = T$ is the measured tension in Newton.

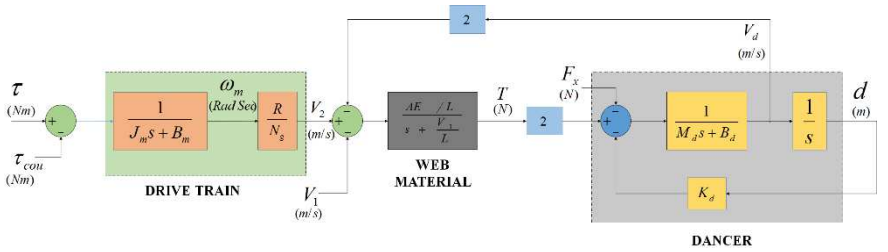


Figure 2
Block diagram representation of WTS with rollers and dancer arrangement

3 Adaptive Model Predictive Controller

One major challenge in maintaining constant web tension is the continuous variation of the web radius due to winding and unwinding operations. The web tension needs to be regulated considering the current radius of the web and generally, sensors are used for this purpose. An accurate determination of web radius, using measurements is rather difficult due to acceleration and de-acceleration of the web and friction in rollers. As a result, the parameter variations in radius cause deterioration in control performance leading to the material loss, downtime, and poor material integrity. In order to overcome this performance loss, variations of the web radius need to be compensated. Furthermore, the performance needs to be optimized to guarantee product quality considering the variation in the parameters and disturbance acting on the WTS. This requires combining adaptation to parameter variations, the prediction on disturbances and optimal control for optimizing the performance in the face of uncertainties. To obtain these two features, this investigation designs an adaptive model predictive controller (AMPC) and an estimator to predict the variations in the parameter. The AMPC design has three components, namely, *(i)* an online model estimator, *(ii)* a prediction model and *(iii)* an online optimization routine. The schematic of AMPC design for WTS is illustrated in .

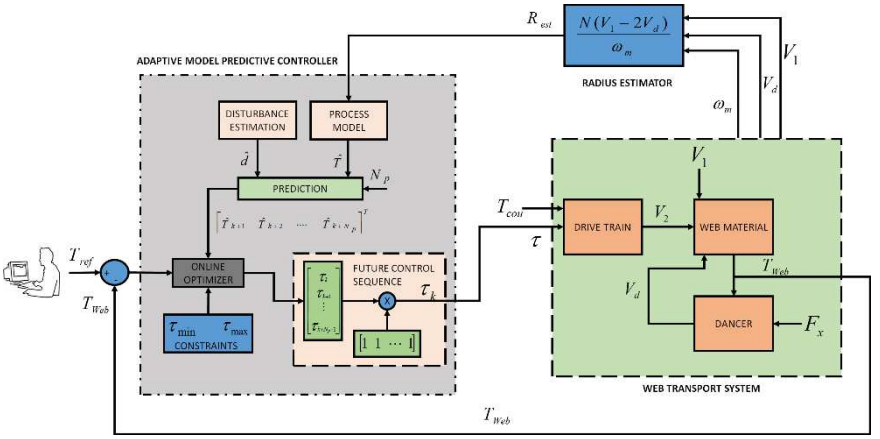


Figure 3
Schematic of Adaptive MPC design for WTS

The online estimator model estimates the web radius from the output data namely, the angular velocity of the motor, un-winder velocity, and dancer velocity, at each time epoch. The estimated radius is given by

$$R_{est} = \frac{N(V_1 - 2V_d)}{\omega_m} \tag{7}$$

The estimated web radius (R_{est}) updates the plant model in (5) during each time period. The prediction model uses this updated model to compute the prediction matrix in (8).

$$P = \begin{bmatrix} CA \\ CA^2 \\ \vdots \\ CA^{N_p} \end{bmatrix} \text{ and } H = \begin{bmatrix} CB & 0 & 0 & \cdots \\ CAB & CB & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ CA^{N_p-1}B & CA^{N_p-2}B & CA^{N_p-3}B & \ldots \end{bmatrix} \tag{8}$$

To obtain optimal control input, an optimization routine (J) is formulated with the variables of interest, the reference tension (T_{ref}), prediction model (P and H), control input (τ) and the disturbance matrix (F). The optimization problem in (9) models the adaptive MPC controller for the prediction horizon and is solved using quadratic programming (by invoking conic solvers) considering the constraints on the WTS.

$$\begin{aligned} &J = \left\| T_{ref} - H\Delta\tau - P\hat{x}_k - Fd_x \right\|_2^2 + \lambda \left\| \Delta\tau \right\|_2^2 \\ &Sub.to. \\ &\tau_{\min} \le 0 \le \tau_{\max} \end{aligned} \tag{9}$$

Solution to the constrained optimization problem provides the control moves for the prediction horizon (N_p.) The first among the control input is applied and the procedure is repeated during each step.

4 AMPC Algorithm for Web Tension Control

This section presents the AMPC algorithm for the web tension controller. The execution sequence of AMPC has six steps as shown in Figure 4.

Step1: Update the parameter and measurement

In the first step of the algorithm, the measurements on web tension measured using load cells are updated. Then, the web radius ($\boldsymbol{R\_est}$) information obtained from the online parameter estimator block updates the plant model during the current time epoch.

Step 2: Model update and construction of prediction matrices

Using the estimated web radius ($\boldsymbol{R\_est}$), the prediction matrices in (8) are updated.

Step 3: Compute control inputs

The reference web tension (T_{ref}), prediction models ($\boldsymbol{P}$ and H), disturbance model (F) and the operating constraints ($\tau_{\max}, \tau_{\min}$) are used to solve the constrained optimization problem in (9) to obtain control inputs (τ).

Step 4: Receding horizon input

The online optimizer provides control inputs (τ) for N_p time steps. Out of N_p control inputs, the first control input is applied and the rest are discarded to implement a receding horizon control.

Step 5: Estimate the parameter

The control input (τ) is applied to WTS, and the web tension ($\boldsymbol{T}_{web}$), un-winder velocity (V_1), dancer velocity (V_d) and angular velocity of the winder (ω_m) are measured using sensors. These measurements are used to estimate the web radius given by (7).

Step 6: Write the output

The web tension ($\boldsymbol{T}_{web}$) and the estimated radius ($\boldsymbol{R\_est}$) are given as the feedback to the controller to update the process model and to calculate the prediction matrices for the next iteration $t+1$.

This execution sequence is repeated for the entire run time. The effects of web radius variations on the web tension, the significance of online estimation model and the performance of AMPC for WTS are discussed in the following section.

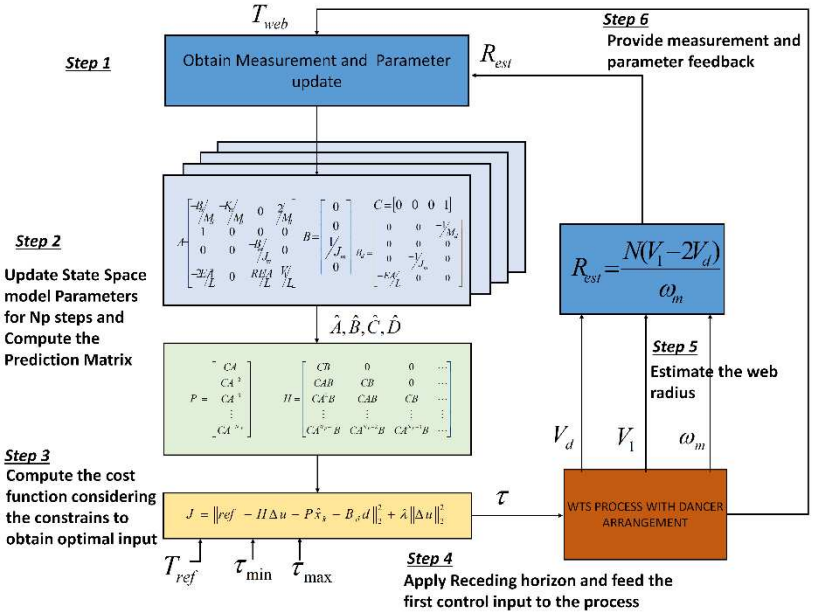


Figure 4
Adaptive MPC algorithm for WTS with parametric variations

5 Results and Discussion

The proposed AMPC is illustrated using simulations with parameters of a web transport system obtained from a paper industry. This section, studies the benefits of adapting the AMPC to control the WTS. To illustrate the benefits of the AMPC, its performance is compared with the conventional MPC used for controlling the WTS presented in Section 2.

During the simulations, the radius of the winder roller is varied from 0.57 to 0.16 m. As the radius changes, the dancer roll moves up or down to compensate the parameter variations. The effect of the radius variations on dancer position is illustrated in Figure 5. It can be observed that as the radius of the winder increases, the dancer roll moves down in the un-winder section to compensate for parameter changes. The estimated web radius obtained using equation (7) and the actual web radius is shown in Figure 6 for the operating range considered in our simulations, and an error of 0.4% was observed in the estimated radius.

The parameters used in our simulations for studying WTS performance are listed in Table 2. The control parameters and the operating constraints for MPC and AMPC controllers are shown in Table 3. These parameters were selected based on the fast dynamics exhibited by WTS and the parameter variations. The prediction horizon should capture the relevant transient information in it to foresee the effect of parameter variations and disturbances. Therefore, based on open loop response characteristics the prediction horizon was selected to be 30 time epochs. Though a control horizon of 10-15 is enough for implementing AMPC, to account for the fast changes in disturbances and parameter variations, a control horizon of 20 epochs was selected. The physical and operating limits used in the simulations are obtained from the process dynamics, these constraints are used in the optimization problem to reflect the actual conditions. The set-points are selected based on the paper variety being processed in the paper industry.

Table 2
Simulation Parameters of WTS model

PARAMETER	VALUE
A	$4.35\text{x}10^{-6}$
E	$4\text{x}10^{9}$
L	0.61
J_m	0.0324
B_m	$0.55\text{x}10^{-3}$
R	$57.3\text{x}10^{-3}$
N	10.1
B_d	500
M_d	6.762
K_d	1131
V_1	0.051
τ_{cou}	$0.0081\text{x}10^{-5}$
F_x	66.3

The changes in web tension for a tension set-point of 70 N using AMPC and MPC controllers are shown in Figure 7. From the result, it is observed that the percentage overshoot is 10.43% with the MPC, as against 5.12% in AMPC. The percentage overshoot reflects the quality loss due to winding operations. Therefore, a reduction of about 50.5% in overshoot that also reflects in the quality of transported material is achieved during the transient stage of the web processing due to AMPC. The improvement is mainly due to the accurate compensation of the web radius in AMPC that in general causes degradation of material quality with the MPC. Second, the settling time of the AMPC is

reduced by 6% leading to faster settling times to ensure required material integrity, thereby reducing material loss significantly. These results illustrate the material savings and quality enhancements achieved using AMPC in process industries.

The improvements achieved using proposed AMPC are illustrated by analyzing performance measures such as, integral square error (IAE), and integral time square error (ITSE). The IAE indicates the improvements in the transient part of the response, whereas the ITSE represents the steady-state improvements. Our results indicate that the proposed AMPC has significant benefits in improving the performance of the WTS by up to 4.2% that reflects the improvements achieved using the proposed AMPC. The improvement is mainly due to the incorporation of parameter estimation model within the AMPC controller.

Table 3

Controller parameters of MPC and Adaptive MPC

PARAMETERS	ADAPTIVE MPC	MPC
Prediction Horizon (N_p)	30	30
Control Horizon (N_c)	20	20
$\tau_{\min}$	-0.01	-0.01
$\tau_{\max}$	0.09	0.09
$\Delta\tau_{\min}$	-0.1	-0.1
$\Delta\tau_{\max}$	0.3	0.3
λ	0.4	0.4

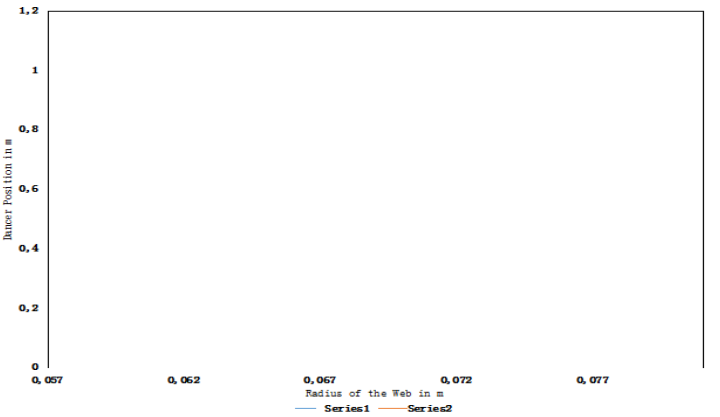


Figure 5

Position of the dancer with respect to the variation in the winder radius

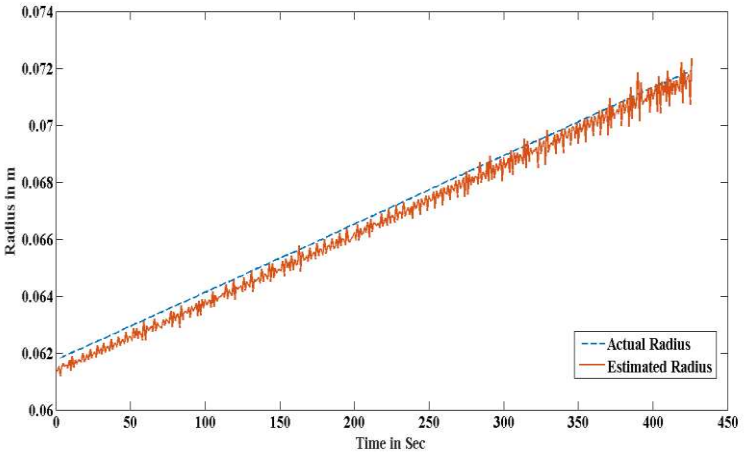


Figure 6

Comparison between actual winder radius and the online estimated radius

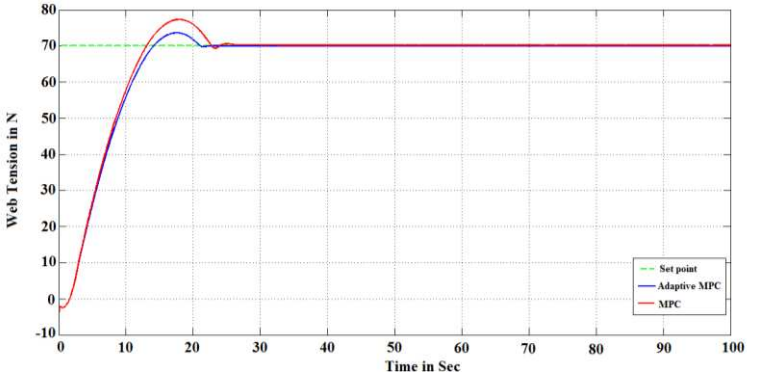


Figure 7

Responses of Adaptive MPC and MPC to a reference web tension of 70 N

Table 4

Transient performance of Adaptive MPC and MPC controller

CONTROLLER	RISE TIME (Sec)	SETTLING TIME (Sec)	PEAK OVERSHOOT (%)
ADAPTIVE MPC	7.75	20.9	5.16
MPC	7.53	22.3	10.43

Table 5

Performance indices of Adaptive MPC and MPC controller

CONTROLLER	IAE	ISE	ITSE
ADAPTIVE MPC	5.19	259.54	723.29
MPC	5.43	256.6	725.95

Conclusion

This investigation proposed an adaptive model predictive controller for web transport systems that combines adaptive, optimal and predictive features. The confluence of these highly desirable features in web transport systems led to improvements in performance, quality and reduced downtime. The main building blocks of the controller are a parameter estimation block, prediction model and an online optimization routine. The parameter estimation block uses the variables form WTS process to estimate the changes in web radius. The web radius estimation obtained is close to the actual radius with an estimation error of 0.4%. Using the estimated web radius, the prediction model is updated at each time epoch to handle the parametric variations. An online optimization routine is used to provide an optimal control input considering the physical and operating constraints. To verify the performance improvement obtained by AMPC, its performance is compared to conventional MPC. Our result shows that AMPC handles the parameter variations effectively with 6% increase in material integrity and 50.5% increase in material quality. Further, the AMPC controller performance in regulating web tension shows improvement up to 4.2% than conventional MPC.

Acknowledgement

This investigation is supported by DST-SERB (SB/FTP/PS -061/2013). Dated 24/02/2014.

References

[1] Baosheng Wang; JianminZuo; Mulan Wang; HongyanHao: Model Reference Adaptive Tension Control of Web Packaging Material, Intelligent Computation Technology and Automation (ICICTA), 2008 International Conference on, Vol.1, No., pp. 395-398, 20-22 Oct. 2008

[2] Boulter.T, ZhiqiangGao:ANovel Approach for On-Line Self-Tuning Web Tension Regulation, Control Applications, 1995, Proceedings of the 4th IEEE Conference on , Vol. 91, No. 98, pp. 28-29, 1995

[3] Hou, Y., Gao, Z., Jiang, F., Boulter, B. T.: Active Disturbance Rejection Control for Web Tension Regulation. Proceedings of the 40th IEEE Conference onDecision and Control, Vol. 5, IEEE, 2001

[4] JeppeSondergaard Larsen, Peter Kai Jenson: Adaptive Control with Self-Tuning for Center-driven Web Winders, Master Thesis, Aalborg University, 2007

[5] Jimoh Pedro, John Ekoru: NARMA-L2 Control of a Nonlinear Half-Car Servo-Hydraulic Vehicle Suspension System, Acta Polytechnica Hungarica, Vol. 10, No. 4, pp. 5-26, 2013

[6] Kim, Jeetae: Development of Hardware Simulator and Controller for Web Transport Process. Journal of Manufacturing Science and Engineering, 128.1, 378-381, 2006

[7] Ku Chin Lin: Frequency-Domain Design of Tension Observers and Feedback Controllers with Compensation, IECON 02 [28th Annual Conference of the Industrial Electronics Society, IEEE 2002], Vol. 2, pp. 1600-1605, 2002

[8] Ku Chin Lin: Observer-based Tension Feedback Control with Friction and Inertia Compensation, IEEE Transactions on Control Systems Technology, Vol. 11, No. 1, pp. 109-118, 2003

[9] Liuping Wang,: Model Predictive Control System Design and Implementation using Matlab, Springer, 2009

[10] Muthukumar. N, Seshadhri Srinivasan, K. Ramkumar, P. Kavitha, and Valentina EmilaBalas: Supervisory GPC and Evolutionary PI Controller for Web Transport Systems, Acta Polytechnica Hungarica, Accepted for publication, 2015

[11] Pagilla P., Dwivedula R. and Siraskar, N.: A Decentralized Model Reference Adaptive Controller for Large-Scale Systems, IEEE/ASME Transactions on Mechatronics, 12, 154-163, 2007

[12] J. A. Rossiter: Model-based Predictive Control - A Practical Approach, CRC press, 2004

[13] Sakamoto, Tetsuzo, and Yoshikazu Fujino: Modelling and Analysis of a Web Tension Control System, Proceedings of the IEEE International Symposium on Industrial Electronics, 1995, ISIE'95, Vol. 1, IEEE, 1995

[14] Valenzuela, M., Bentley, J. M., Lorenz, R. D.: Sensorless Tension Control in Paper Machines, Conference Record of the 2002 Annual Pulp and Paper Industry Technical Conference, Vol., No., pp. 44,53, 17-21 June 2002

[15] Weixuan Liu, Davison, E. J.: Servomechanism Controller Design of Web Handling Systems, IEEE Transactions on Control Systems Technology, Vol. 11, No. 4, pp. 555-564, July 2003

[16] D. P. D. Whitworth, M. C. Harrison: Tension Variations in Pliable Material in Production Machinery, Applied Mathematical Modelling, Volume 7, Issue 3, pp. 189-196, June 1983

[17] Xiong, T., Cai, W., Xiong, Y. & Zhang, R.: Dynamic Matrix Control of the Lateral Position of a Moving Web, International Conference on Mechatronics and Automation (ICMA), 2012, 1091-1096, 2012

[18] Young G. E., Reid K. N.: Lateral and Longitudinal Dynamic Behavior and Control of Moving Webs, Journal of Dynamic Systems, Measurement, and Control, Vol. 115(2B), pp. 309-317, 1993

[19] Boulter. T, Zhiqiang Gao: A Novel Approach for On-Line Self-Tuning Web Tension Regulation, Control Applications, 1995, Proceedings of the 4th IEEE Conference on , Vol. 91, No. 98, pp. 28-29, 1995

Emergency Situations Management with the Support of Smart Metering

Judith Pálfi, Peter Holcsik

Óbuda University,
Research Group of Applied Disciplines and Technologies in Energetics,
Bécsi út 96, 1034 Budapest, Hungary
palfi.judith@kvk.uni-obuda.hu, peter.holcsik@elmu.hu

Today's extreme weather conditions cause more and more emergency situations every year. These emergencies represent a big challenge for today's power supply companies.
The task to be solved in these critical situations is the processing of the huge volume of incoming data and their transmission by one-line messages (tolerant protection signals) to the operation controllers. In addition to this, high-tech equipment is to be provided in order to handle the large quantity of internal and external data and information.
According to the opinion of the Research Group of Applied Disciplines and Technologies in Energetics (AD&TE), the optimal solution to deal with the above challenges is the use of Smart Metering devices. The implementation of these devices will result in two directional communication which reduces troubleshooting time and simultaneously supplies the required quantity of information.

Keywords: management of emergency events; Smart Metering

1 Introduction

'Climate change represents a big challenge for human societies. Anthropogenic activities (air pollution – with a growing amount of greenhouse gases, environmental damages, land overuse and overpopulation) contribute to the climate change both directly and indirectly. Our own health defense system is weakened by these irresponsible activities. A multitude of emergency events and catastrophes illustrate the way the climate conditions become more and more extreme. The relation between global climate change and extreme weather events is obvious - any climate change causes changes in the weather conditions.' [1]

Extreme weather conditions have an important influence on the activities of network operators as well. Extremely high temperatures can damage the underground power cables, strong winds and ice formation can lead to the failure of the overhead electrical transmission lines disrupting the continuity of energy supply. [11] [12] [13] [14]

Todays difficult financial situation of utility companies causes the decline of the operation efficiency and reduces maintenance investments [15] [16]. The lack of investments results in a greater density of weak points in the networks, increasing the risks during their operation.

Figure 1
Overhead lines affected by an emergency event in the area of ELMŰ-ÉMÁSZ
(ELMŰ Hálózati Kft., 12. 2014.)

2 Electric Supply Companies Tasks during Emergency Events

The electric supply companies tasks during emergency events can be divided into the following: general preparation, direct preparation, emergency management and evaluation. These tasks are handled by the different departments of the Distribution System Operators (DSO) (Figure 2)

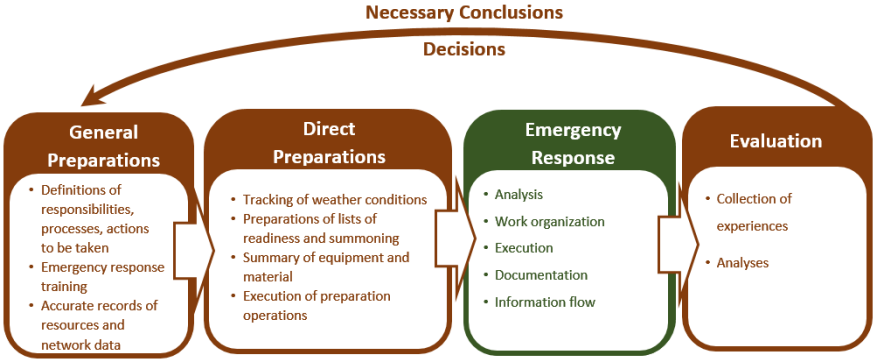


Figure 2
Task list for preparation to emergency events and restoration of services [3]

The general preparation for emergency events happens during normal operational conditions, well before the emergency event takes place. During this preparatory period, procedures and responsibilities are defined, tasks and duties identified.

Comprehensive resource- and equipment registers support the organisation of the work of specialists dealing with emergencies. The emergency response cycles are worked out in detail, the responsible people to make decisions are designated (Figure 3).

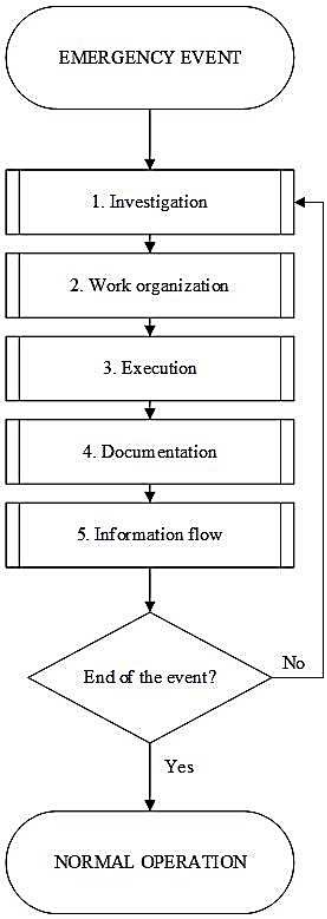


Figure 3

Flow chart for failure management in the power supply network during emergency events

Emergency events initiate complex processes. The first sign of an emergency event (inside the operation and control systems) is given by a sudden surge in the number of protection signals and consumer error reports. These together trigger the situation analysis and the delimiting of the affected geographical areas. Operation and work controllers take action, evaluate the available resources and, whenever the case, order the deployment of additional personnel and equipment.

The organization of troubleshooting activities is carried out taking into account the professional competences of the teams, the qualification level of the experts, the geographical position of the equipment and errors requiring attention. Due to efficiency reasons, the same specific teams handle the errors geographically close to each other or which suggest similar type of failures (for example: reports about flooding of a power station which requires pumps). During normal operational conditions, the distribution of the error or failure reports (failure addresses) is managed by the mWFM (automatic dispatcher) [8]. However, the mWFM is not capable of managing emergency events.

A new work organization strategy could be to reverse the above workflow by the distribution of a whole region or area to a working team - geographical area based work assignment instead of a failure report based one. In this case, the documentation of the management of error addresses, failure reports and protection signals will be handled post-event.

Following receiving the work instructions (nowadays in a digitalized form) the troubleshooting teams are dispatched on the spot and proceed to the delimitation of the exact position of the failure. Only after localizing and identifying the type of error – failure of an equipment or network element - can they proceed to carry out the troubleshooting works. On MV networks this working process is directed by the MV operation control, on LV networks by the LV operation control. The latest coordinates also the troubleshooting works.

Documenting the process is an important part of the troubleshooting. During normal operations this is relatively easy to execute, however during emergency events it can be a very demanding and complicated task. Reported back information is processed continuously which may initiate further processes through which the implementation and development of additional equipment and material, additional personnel or resources, external subcontractors or organizations (e.g. disaster management) may be required.

During emergency situations troubleshooting teams change continuously their geographical position. The materials, equipment supply and stock varies. Accordingly, the teams, equipment and stocks can be reorganized and optimized. Reorganisation of troubleshooting teams takes into account the peronnel competence, qualifications, exhaustion degree, etc. The incoming of failure reports, the development and “discovery” of new error locations lasts until the very end of the state of emergency. The emergency response cycle (Figure no. 3) can be as short as a couple of minutes during an intensive storm.

Continuous internal and external communication of the actual situation is an organic part of the process.

3 LV Operation Control Support during Emergency Events at Present

During an emergency the top priority task is to support the troubleshooting team(s) in charge to restore the normal operations mode. The technical staff (regional managers, controllers, directors, etc.) of the DSO helps in the localization of the failed network elements, in the unification of the interrelated error addresses, in the organization, troubleshooting and data sharing.

However, in many cases, due to the large amount of error reports, the LV operation control dispatchers fail to properly manage the emergency event.

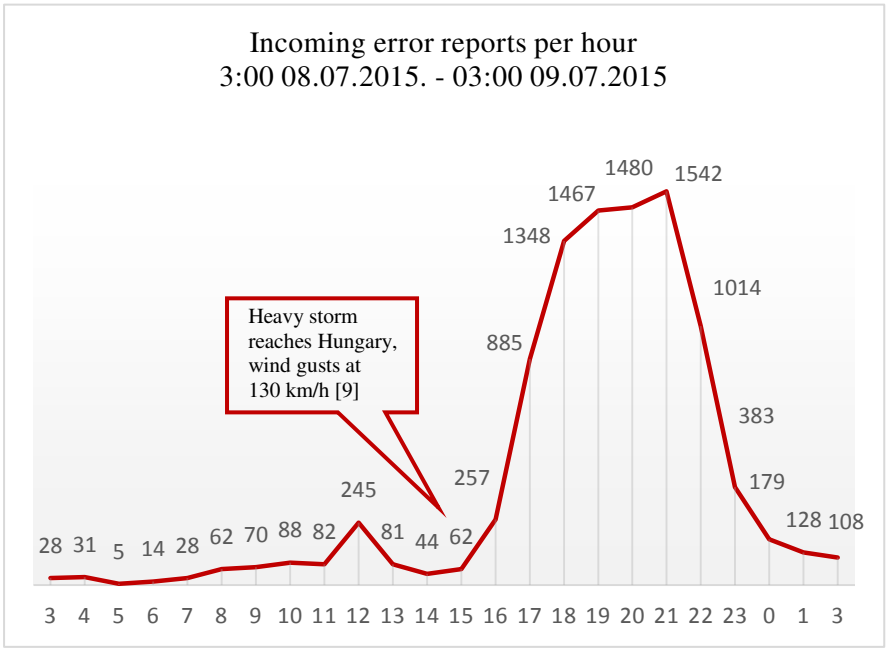


Figure 4
Incoming error reports (per hour) ELMŰ-ÉMÁSZ,08–09. 07. 2015.

This big volume of data hinders the optimal decision making process. Temporary, on short-term, the decision making process can be paralyzed. On longer terms, with regards to the entire time span of the event, the troubleshooting process will be slowed down.

Operation control systems currently filter and summarize error reports per LV circuits (provided that there is available network data in the system to the report address). In a given district of a transformer outage, relations to MV errors are handled by the previously mentioned human resources. This task can be replaced in the near future by proper digital support systems. The workflow made up by the

processing of the huge amount of incoming data (Big Data) and their transmission without redundancy to the troubleshooting teams in the form of one-line messages (tolerant signal) can be automated.

Another task is to provide the large amount of information requested by the different parties: authorities responsible for the protection against catastrophes, media (newspapers, TV, etc.) and internal departments (executives, fellow departments, etc.) This is a top priority duty as failure to perform it properly can cause a decreasing of the public satisfaction (due to itspower dependence': heating and cooling, traffic, communications, IT, cash registers, healthcare equipment, internet, etc.), negative media attitude and political uncertainty among others. [4] [21]

The management of these problems can be optimized by using Smart Metering. The implementation of this tool creates a two-way flow of communication which significantly improves the emergency event management and responds to the above-mentioned information demands.

4 Emergency Events Management with Smart Metering

The main goal of the Smart Metering (SM) is to improve the efficiency of the given energy systems (these can be: electricity, gas, water, central heating, etc.) and to increase the quality of their service. In order to control the demand, to implement intelligent metering and influence the energy consumption habits, the energy supplier needs to better understand the consumer. Using this knowledge he will be able to offer customized solutions. This will allow offering discounts to the consumers that will improve the utilization and optimisation of the capacities of his own systems as well. For example, the energy supplier may offer a price discount for consumptions outside of the peak time [5].

Further advantages of implementing Smart Metering are:

- the increase in energy efficiency by rending transparent and trackable the consumers' energy consumption,
- the possibility of monthly invoicing based on real consumption (instead of lump sum rates and 'monthly reporting') by connecting Smart Meters to the DSO's invoicing systems [5],
- replacement of HFKV and RF methods [19] [20],
- fast and reliable data service for consumers using the two-way communication via the SM display: information about the start of repair works, notice before the start of maintenance works, 'calm down messages' during emergency events, etc. (the precondition is that the Smart Meters will be equipped with battery and 'acknowledged' button),

- more transparent tracking of produced and consumed renewable energy,
- tracking down the ilegal consumption [6] [18].

In the event of a failure at 0.4 kV LV or at MV without substation protection signal (three-phase breakdown, meltdown of the LV/MV transformer primary fuse, heat protection meltdown, etc.) the operation control gets instant, full-scale report about the consumers without service, the failure locations within the network and their number. By this, the delimitation and the repair of errors can be more quickly assigned and executed. At present, the information about these types of failures is received only via consumer reports. [6]

A further problem to solve is the synchronization of the addresses, that is the assignment of consumer addresses to the network elements. At present, this type of data is not fully available on the LV network.

The error reports are organized on the lines of a monitor of the coordinating dispatcher:

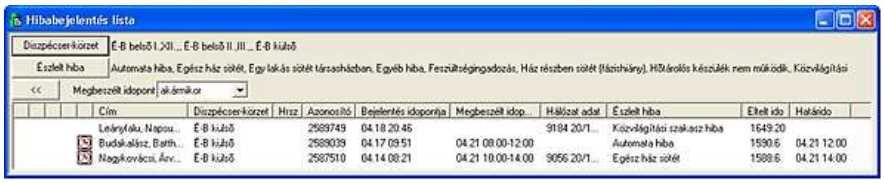


Figure 5

Error reports module of the LV operation control and work management system 'Mirtusz' at ELMŰ-ÉMÁSZ [22]

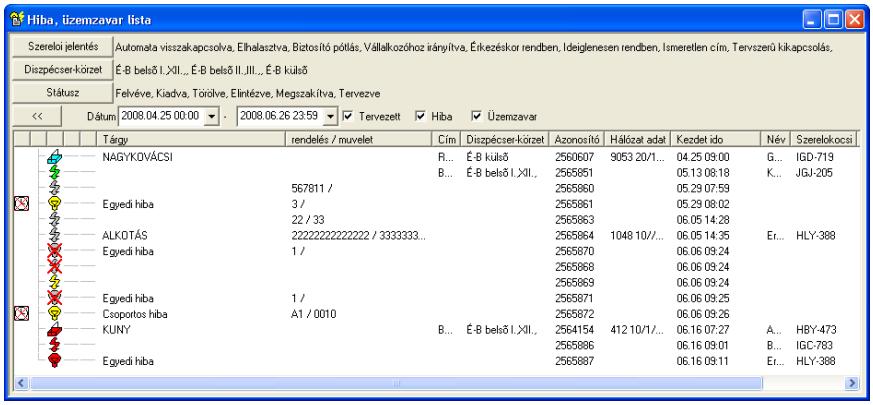


Figure 6

Error and malfunction list module of the LV operation control and work management system 'Mirtusz' at ELMŰ-ÉMÁSZ [22]

During normal operations the magnitude of 10-50-100 reports is easily understood and managed. However, this type of visual presentation is insufficient during an

emergency event. The solution to the problem is given by ***the introduction of tolerant signals*** based on SM, the ***identification of affected network elements*** and the ***modern visual presentation of data and information.***

There are currently available technologies for the visual presentation of LV and MV network data (outages, error reports, etc.). However, these are not used in real live situations. The input data for the E-software currently consists of error reports received via Tele Centers in the form of error addresses (Figure no. 7).

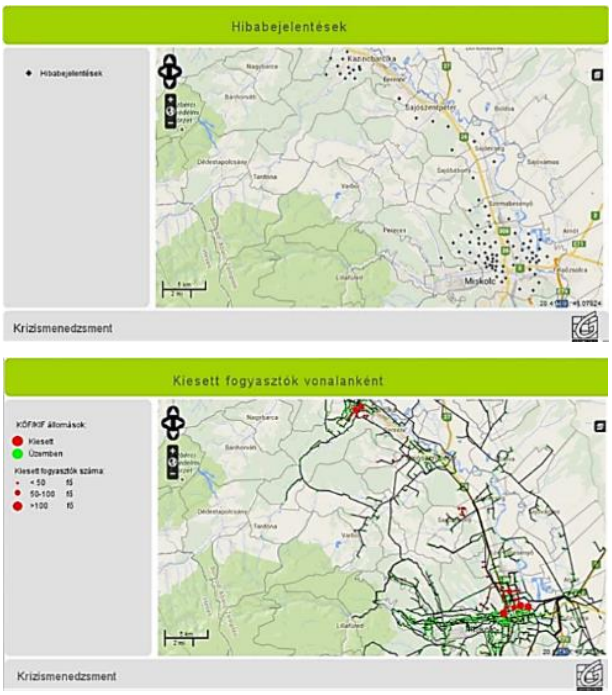


Figure 7
Displays for visual presentation of error addresses (left) and MV outages (right), developed by Geometria kft. [3]

The on-line and SM-supported displays applied in LV control centers provide significant help during normal operations. However, their use proves to be especially advantageous during emergency events.

In addition to presenting error addresses and failed LV network elements, the equipment is required to manage the on-line presentation of

- the position of the troubleshooting teams (with detailed features list displayed as pictograms),
- the network elements affected by HV/MV malfunctions (which, because of the network hierarchy, also affect the LV network elements),

- the network elements that are affected by the planned works (this gains function primarily during normal operations, since planned switchings are postponed in large quantities during emergency events).

A further advantage of the system is given by the high quality sharing of information with the external and internal data demanders: graphic, highly accurate and online data is available to the control, support personnel, executives and external organizations (e.g. media, authorities for protection against catastrophes, etc.).

5 Real Time External and Internal Communication with SM

Outage data can be used online. It can be made public in real-time providing instant and detailed information to the public and other organizations about the state of the troubleshooting process. In practice, the external-oriented information would be published based on the data of SM by the person responsible for the internal communication of the energy supply company:

- 'Informative-calming' messages to the consumers affected by outages (per region or per district) received on the display of the measuring meter.
- Situation reports sent to external organizations and authorities responsible with the protection against catastrophs.
- Information to the media (in the form of situation reports).
- Internal information for the executives, in the form of situation reports, for the preparation of further measures and for decision support.

The number of incoming calls of Tele Centers (TC) can be reduced significantly by the implementation of new communication systems.

By implementing the real time communication with SM:

- Accurate and detailed information about the emergency event can be provided to all the parties involved in the troubleshhoting process.
- Disaster control and external organizations can optimize the coordination of equipment (aggregators, pumps, special vehicles…) and staff more efficiently.
- Media receives accurate information.
- The more accurate information for the executives will help taking optimal, precise and well-grounded decisions.

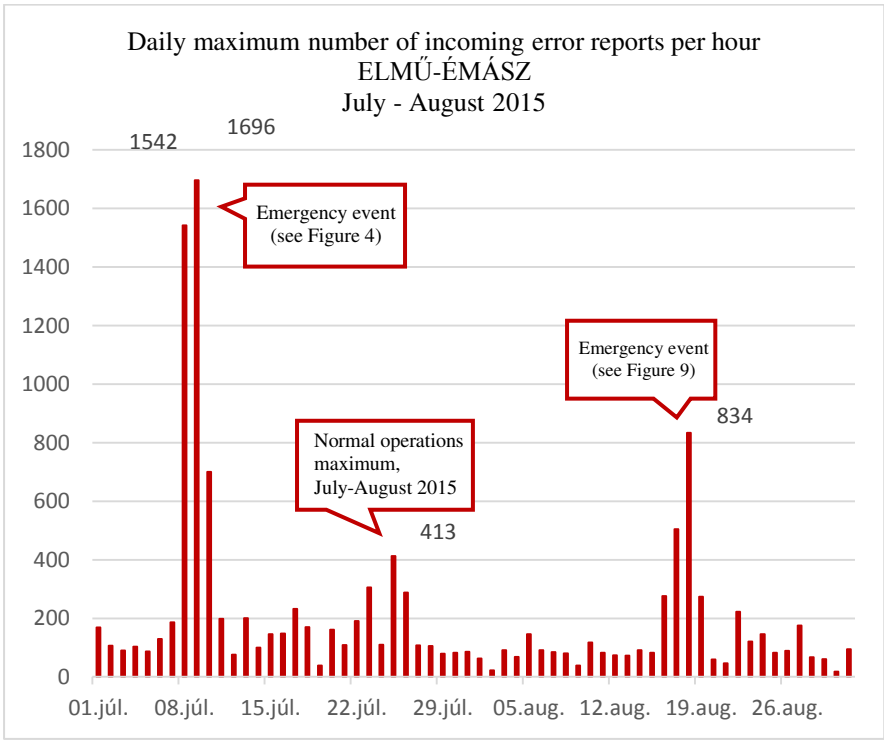


Figure 8
Daily maximum number of incoming calls (per hour) during normal operations and emergency events, ELMŰ-ÉMÁSZ, July-August 2015

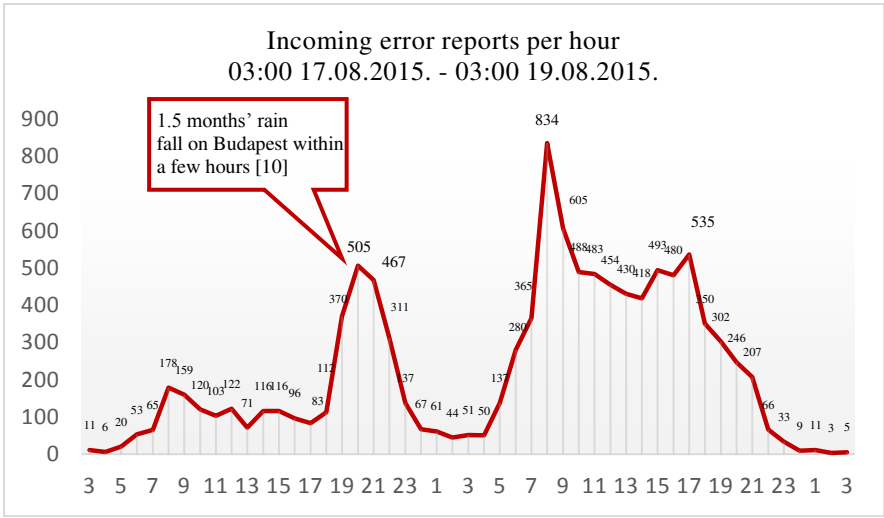


Figure 9
Incoming error reports (per hour), ELMŰ-ÉMÁSZ, 17-19. 08. 2015

Conclusions

Electric networks can suffer serious damages during extreme weather conditions. The rapid restoration and reparation is a complicated task. The emergency teams are able to fulfill their job with the help of current operational systems. However, in order to mitigate damages and to shorten service outages caused by the deficient operation of the power system new developments at system level are required. As an example, the establishment of optimized LV operation control centers and the development of technologies and tools which can provide quality support to the troubelshooting process. Further research of these support systems and the implementation of the findings into the real life practice will result in a higher quality of service.

References

[1] László Teknős: System of Protection against Emergency Events Caused by Extreme Weather Conditions and Civil Defense Tasks in Outages Areas, Hungarian Academy Science, 2013, p. 1

[2] István Molnár Head of System Operations: Preparations for Emergency Events at E.On Distributors, E.O.N., MEE Itinerant Congress, 2015, p. 4, (http://www.mee.hu/files/files/b5_molnari.pdf) ISBN: 978-963-9299-27-6

[3] Tibor Tenke, dr. István Cseke: IT Support for Critical Failure Situations, Geometria kft., MEE 62. Itinerant Congress, Siófok, 2015, p. 15, (http://www.mee.hu/files/files/b5_tenket.pdf), ISBN: 978-963-9299-27-6

[4] Tibor Tenke, dr. István Cseke, IT Support for Critical Failure Situations, Geometria ltd., MEE 62. Itinerant Congress, Siófok, 2015, p. 4, (http://www.mee.hu/files/files/b5_tenket.pdf), ISBN: 978-963-9299-27-6

[5] István Szabó M.: This will be a Bigger Business than the Internet, 18. 04. 2014, (http://hvg.hu/gazdasag/20140418_Nagyobb_uzlet_lesz_ez_mint_az_internet)

[6] Attila Fábián, dr. György Morva: What about you, Smart Metering?, VL magazine, 2014, (http://www.villanyszaklap.hu/lapszamok/2014/aprilis/2993-2014-04-18-11-20-39)

[7] István Molnár, Head of System Operations: Preparations for Emergency Events at E.On Distributors, E.ON, MEE Itinerant Congress, 2015 (http://www.mee.hu/files/files/b5_molnari.pdf), ISBN: 978-963-9299-27-6

[8] ELMŰ-ÉMÁSZ MIRTUSZ Work Control System 3.0: Mobile application and automated scheduler, Geometria ltd., 2015, (www.geometria.hu/?p=3079)

[9] Storm Struck the Country, 2015, (http://www.hirado.hu/2015/07/08/megerkezett-a-vihar-a-dunantulon-mar-elsofoku-a-riasztas/)

[10] Ferenc Bakró-Nagy: One and a Half Months' Rain Torrents on Capital, 2015 (http://index.hu/video/2015/08/17/viz_ala_kerult_budapest_eso/)

[11] Disaster Control's Announcement: About The Last 24 Hours of Extreme Weather, 2013 (http://www.katasztrofavedelem.hu/index2.php?pageid=szervezet_hirek&hirid=1634)

[12] Freezing Rain: Roads Still Closed in Budapest, Weather Report, 2014 (http://www.ma.hu/idojaras/232895/Onos_eso_Budapesten_meg_tobb_ut_van_lezarva)

[13] Ice Situation: Restoration Can Take Months, 2014 (http://nol.hu/mozaik/jeghelyzet-14-ut-van-lezarva-Budapesten-1502023)

[14] Judith Füzes, Ádám Draveczki-Ury, Gábor Kálmánfi, László Sályi Gergő, Szűcs, MTI; Tibor Benkő: The Army Meets All Incoming Requests, As Has Done So Far, 2013 (http://www.honvedelem.hu/nyomtat/36948)

[15] Elmű Pocket Guide, ELMŰ ltd., 2015 (http://kismarosikikialto.hu/elmu-kisokos/)

[16] dr. Norbert Boross: Thousands Lose Jobs Because of Battle for Utility Cost Reduction, 06. 02. 2013. 11:31, Last accessed: 05. 11. 2015. 22:14, http://www.atv.hu/belfold/20130205_boross_norbert

[17] Tibor Tenke, dr. István Cseke: IT Support for Critical Failure Situations, Geometria ltd., MEE 62. Itinerant Congress, Siófok, 2015, pp 4 (http://www.mee.hu/files/files/b5_tenket.pdf) ISBN: 978-963-9299-27-6

[18] Roland Menyes, Ágoston Salacz, Ervin Szabó: System, System Description version 2011.09.13, Prolan Energy Management, 2011, p. 17 (http://www.prolan.hu/wp-content/uploads/2011/09/Energiamenedzsment_Rendszerleiras_SD_HU.pdf)

[19] dr. Dávid Raisz Dániel Divényi: Intelligent Metering and Influence of Consumers', digital note, 2015 (ftp://ftp.energia.bme.hu/pub/Energiaellatas%20es%20-gazdalkodas%20-%20B/Fogy_Vez_SM_2015.pdf)

[20] dr. Bálint Kiss, Takács Tibor, dr. Gábor Vámos, Zsolt Béla Gombás, Mihály Gábor Péter, Ferenc Szűcs, Imre Veisz: Solutions for Direct and Indirect Control Options in Smart Metering, Algorithms for Analysing T-Curves and for Controlling, MEE 57. Itinerant Congress – Conference and Exhibition, Siófok, 2010, (http://www.mee.hu/files/images/5/MEE57_bme-eon-mee-a3_4_v3_end.pdf)

[21] Report In Progress About Experiences of Extraordinary Snow, 2013 (http://tuzoltosag.info/cikk/20130317_jelentes_keszul_a_rendkivuli_havazas_tapasztalatairol/)

[22] MIRTUSZ Work Control System, User Manual, ELMŰ ltd. and ÉMÁSZ ltd.

Human Resource (HR) Outsourcing in European Compensation Management in the Light of CRANET Research

Nemanja Berber, Agneš Slavić

University of Novi Sad, Faculty of Economics in Subotica, Segedinski put 9-11, 24000 Subotica, Republic of Serbia, berber@ef.uns.ac.rs, slavica@ef.uns.ac.rs

Abstract: The purpose of this paper is to explore the practice of outsourcing in Human Resource Management (HRM) in countries of Europe-EU and Serbia. An analytical exploration of available literature in the area of outsourcing was made, as well as a statistical analysis of the CRANET research data to determine the actual level of outsourcing in HRM in companies around the world, and to make comparison between Serbian compensation outsourcing practice and countries of 20 EU countries. Results of the analysis pointed out that outsourcing of payroll and benefits is used lesser than outsourcing in the field of pensions. Serbian companies use outsourcing at a new low level. There are statistically significant positive correlations between outsourcing of payroll, pension and benefits among each other, and with the number of employees in HR department. A multiple regression model was used to explore the predictors of HR outsourcing. The differences between EU countries in the area of compensation outsourcing have been discussed.

Keywords: HR; Human Resource; outsourcing; compensation; EU; Serbia; Cranet

1 Introduction

In today's turbulent economic and political environment modern organizations are searching for possibilities to ensure competiveness, as well as, sustainable and long term development. The last economic crisis, followed by political, cultural, social and moral crisis, made new conditions for organizations, and many of them could not survive. This is even more important in the context of international business and all the issues arising from it [13, 21]. Since there is a higher link to foreign consumers, competitors and suppliers [22] there is also a higher sensitiveness of organizations regarding changes on international market. The improvement of the quality of business processes is the fact that enables higher competitiveness [25]. In order to survive and develop business, management and organization gain special importance especially when we talk about intellectual capital, where human capital is one of the most important [35, 51].

Human resource management includes different possibilities and activities for successful human capital management in organizations. There are numerous researches made in the recent years related to the themes which are important for the HRM development, such as strategic human resource management [14, 26, 50, 51], human resource information system [31, 39, 45, 46], human resource outsourcing [1, 7, 10, 20, 23, 24, 44, 47], compensation [9, 20, 42, 43], CSR and HRM [6], etc. Human resource outsourcing gains special importance in this research.

The usage of the external providers, or outsourcing, is one of the ways that can be helpful in achieving business success in the field of HRM. Namely, HRM represents a set of general and specific activities, aimed to assure, maintain and manage people in one organization and it is recognized as the factors that contribute to the competitiveness of organizations [2, 5, 6, 8]. Interest in this kind of practice was presented in researches from EU countries [1, 10, 23], Australia [47], Hong Kong [16], Taiwan [41], Canada [48], USA [29], etc.

The main goal of this paper is to present the practice of HR outsourcing (HRO) in the area of compensation in countries of Europe, with special regard to EU countries in the comparison with the Republic of Serbia. Authors made statistical analysis of the data collected in the research period from 2008 until 2010, under international CRANET project. Special attention was dedicated to the area of external providers and outsourcing practices in area of compensation inside the HRM. Information was captured for countries of EU region and Serbia, and a comparison has been made. It was interesting to see the practice of HRO in these two regions, especially in the light of Serbian process of approaching and future accession to the European Union. Statistical analysis was performed using the SPSS program. Statistical techniques, including the descriptive statistics, Spearman rho correlation and a multiple regression model have been used.

The paper consists of three parts. In the first part, the authors presented basic assumptions on outsourcing, its advantages and disadvantages, research from the past related to the HR outsourcing. The second part of the paper is dedicated to the presentation of the methodology and data used for the analysis. The authors presented CRANET project and main dependent and independent variables for regression model. The third and final part of the paper includes the summary of all results of the theoretical and empirical analysis, as well as the discussion of the main differences between EU countries and Serbia in area of compensation outsourcing. The empirical data used in the present research, pointed to the actual practice of HRO in the area of compensation and to the organizational predictors of the usage of HRO. This paper adds new value to the concept of HRO since this area of HRM is insufficiently explored.

2 Theoretical Background

Outsourcing is considered an old business method [3]. It peaked in the 1970s, when, as stated by Kakabadse and Kakabadse (2000), large and diverse corporations were considered to be underperforming. More pronunciation of outsourcing came in the early 1980s with the onset of global recession. Outsourcing became an important business approach and accordingly, a competitive advantage may be gained if products or services are produced more effectively and efficiently by outside suppliers [32, 52]. It also gains flexibility and core stability by focusing on the core elements of the firm, and other factors of improvement of the firm. Some of these elements included cost reduction, managing a high number of employees, etc. This action can be applied to both requirements for components and business services (which include HR) [3].

HR outsourcing is defined as placing responsibility for various elements of the HR function with a third-party provider [49]. In one way, HR outsourcing is seen as an instrument of creating time for HR to become a strategic partner, and in another way, as a cost cutting instrument, gradually reducing HR staff [19]. Outsourcing allows firms to focus on their core competences by relocating limited resources to strengthen their core product or service [30] and to strategically use outside vendors to perform service activities that traditionally have been internal functions [15, 38].

The typical reasons for outsourcing include seeking specialist services and expertise, cost reduction, and enabling HR specialists to focus on strategic role. Outsourcing of the HR activities to another company will not only reduce the costs of the company, but will also increase the possibilities for investment in the core elements of the business. HRO decisions are frequently a response to an overwhelming demand for reduced costs for HR services. The costs that were intended for the elements that were considered noncore are lessened. They include very important ones, such as regular salary, to those essential to the HR, such as training and other needs for employees. In the research of Susomrith and Brown (2013) three common reasons for outsourcing of HR functions were underlined: to acquire specialized HR capabilities, to improve quality and efficiency, and to free resources to concentrate on the strategic role of HR. Besides, some reasons for HRO are improving productivity, flexibility, speed and innovation in developing business applications, access to new technologies and skills, transform organization, increase service value, etc. [17, 40]. The benefits and arguments *for and against* HRO authors Cooke, Shen and McBride have discussed in their theoretical research in 2005 (Table 1).

Table 1
Perceived Benefits and Potential Adverse Consequences of Outsourcing

Perceived Benefits	**Potential Consequences**
Concentration on in-house expertise	Discontinuity of skill supply
Specialist supplier's economies of scale	Loss of in-house knowledge and capacity
Numerical flexibility	Reduction in quality
Shift burden of risk	Higher total cost
Competitive tendering process	Loss of employee morale
Organizational learning from specialist provider	Loss of long-term competitiveness

Source: Cooke, Shen and McBride (2005)

Although Cooke, Shen and McBride (2005) found that 97% of organizations use external providers for at least one HR function, outsourcing is still considered "handle with caution". For example, in Germany many firms have never explicitly considered outsourcing of HR functions. HR outsourcing includes broad range of internal HR functions and the respective, externally procured personnel services such as temporary agency work, payroll accounting, interim management, outplacement services, HR consulting, placement services [1]. In contrast, in Hong Kong, although respondents were generally favorable towards outsourcing, in practice its adoption and diffusion were in a nascent stage [16]. In the article on the development in human resource outsourcing (HRO) in recent years, particularly in the light of the economic recession prevailing since 2007, authors established that companies are increasingly outsourcing a routine HR processes, but in some cases also the critical HR processes are in view to cut costs. But, the same authors stress that while such a strategy could be viable in the short term, its long-term strategic effectiveness is questionable [7].

One more interesting theoretical issue is to determine, mostly outsourced, HR activities. Braun, Pull, Alewell, Störmer, and Thommes (2011) reported the existence of common outsourcing practice: among 1021 firms interviewed, 61.7% buy training services from an external service provider; 54.3% use external legal advice; 49.8% buy services in temporary agency work; 33.6% ask for consulting services and 31.8% use the assistance of external service providers in headhunting. A smaller share of firms procure payroll accounting, placement services, recruitment support, outplacement and interim management externally, or even outsource the complete HR function and purchase all personnel functions externally. Presenting research results from Belgium, Cooke, Shen and McBride (2005) emphasized training and development, staffing, payroll and benefits administration as functions for outsourcing, while Delmotte and Sels (2008) emphasized that 71.8% of organization outsource payroll. Hungarian organizations outsource several HR activities, and around 58% of them use outsourcing for payroll [33]. In Australia recruitment and selection, training, occupational health and safety, payroll and employee benefits have been found as

the top five outsourced HR functions [47]. Since there are continuing pressures to improve administrative efficiency in human resource management (HRM), both the professional and academic literature propose "payroll" as an ideal candidate for outsourcing in order to drive costs down. While key payroll activities were more costly when outsourced, there were efficiency gains in supplementary activities and lesser investment in IT software and maintenance [20]. According to a Greek author, the human resource services that are outsourced can fall into one of the following four categories: recruitment and selection, training and development, pay and benefits, and merger-outplacement-downsizing [34]. Besides, one more interesting research was conducted in Greece examining the effects of company internationalization on the practice of HRM outsourcing: foreign multinationals (MNCs) will use this practice more than native companies [23]. When speaking about MNCs, it is important to mention the research of Poór et al. (2015) on the development of HRM in subsidiaries of MNCs in CEE region. They found that the HRO was mostly used for training/development and recruitment, in periods, 2008/2009 and 2012/2013. The findings of the 2000 CRANET research showed that HRM outsourcing is used to a lesser extent in Greece than other Western economies and that MNCs outsource more HRM services than Greek companies. Authors Štangl Šušnjar, Slavić and Berber (2013) explored which HR activities are outsourced the most. Those were training and development, HRIS and recruitment. Also, the mentioned authors explored differences between those companies that have HR department and those that do not have regarding the usage of HRO. It has been found that companies in the CEE region without HR department, in the case of several HR activities use external providers more than those companies, which have established a separate HRM department. Authors found statistically significant differences ($p<0.05$) in t-test. According to theory and past research results, compensation is one of the most often outsourced HR function. According to Belcourt (2006) HR functions which can be outsourced in the area of compensation are: payroll, benefits, compensation administration, and pension.

Since compensation is one of the most outsourced HR activities, the authors explored empirical data on HRO in this area.

3 Methodology and Sample

The main goals of this paper were to explain the concept of outsourcing in HRM, as well as the practice of HR outsourcing in countries of Europe, with special regard to the EU in comparison with the Republic of Serbia. Authors made statistical analysis of the data collected for the research period, 2008 through 2010, under the International CRANET project. This international organization under the management of the Cranfield School of Management organizes

comparative researches on the policies and practices of human resource management, by using a standard questionnaire. The survey is undertaken approximately every four years which is important for achieving specific kinds of results, in particular country-comparative longitudinal analyses [11, 12]. The purposes of the survey are to provide high quality data for academics, for public and private sector organizations, as well as for students of the field, to inform research and to create new knowledge about human resource management across the world. Despite the limitations of the survey methods, and the methodological constrains, the Cranet network's surveys are providing large-scale empirical data since 1990. Doing so, contributing meaningfully both to the description and understanding of the developments of HRM practices in a continuously growing number of countries and to the theoretical developments in comparative HRM [28]. The questionnaire is divided into six sections:

- Section I: HRM activity in the organization
- Section II: Staffing practices
- Section III: Employee development
- Section IV: Compensation and benefits
- Section V: Employee relations and communication
- Section VI: Organizational details

The questionnaire contained closed questions and respondents were requested to make their choice from sets of alternative, pre-formulated answers largely covering the specific areas of HRM to be studied. The research data was processed by using SPSS and MS EXCEL programs. Special attention was dedicated to the area of external providers and outsourcing practices in wide area of HRM. Information was captured for countries of European Union and Serbia, and a comparative analysis between EU countries (n=3795 companies) and Serbia (n=50 companies) has been made in area of HRO. It will be interesting to see the practice of HRM in these two regions, especially in the light of Serbian process of approaching and future accession to the European Union.

Main research goals proposed in this paper were:

- To identify the level of the usage of external providers for elements of compensation – payroll, pension and benefits in the EU and Serbia.
- To compare and analyze differences between usage of external providers for compensation elements in the EU and Serbia.
- To explore relations between outsourcing of payroll, pension and benefits, and number of employees in the organization, as well as the existence of HR department.
- To explore the predictors of HR outsourcing of payroll, pension and benefits.

The focus of the comparison and analysis is to find the similarities and differences of HRM practices in 20 countries of EU compared to the Serbian findings. Comparison between Serbia and EU was made through descriptive statistic technique. Descriptive statistics was used to explore the level of usage of HRO in several HRM activities related to the compensation system (pensions, payroll and benefits) and to identify those practices that are outsourced the most. In the CRANET research the outsourcing practice was measured by a five-level scale from 0=not outsourced, to 4=outsourced completely. In addition, we used Spearman rho correlation to identify links between HRO for compensation and number of employees in organizations. This test was used since it is the most common non-parametric measure used when data are not normally distributed, as in this case. Spearman's is a non-parametric equivalent of Pearson's correlation that can show whether and how strongly pairs of variables are related. The authors' idea was to explore whether there are correlations between variables related to the outsourcing practice in the area of compensation (interval variables). Besides, a multiple regression model has been used to explore the predictors of HR outsourcing. As predictors we used: number of employees in the organization, number of employees in the HR sector (the log of these variables), sector of business (private or public), industry (service or production), HRO for pension, payroll and benefits, the existence of HR strategy (written, unwritten and no strategy), and the usage of HRIS for payroll (yes or no).

In Table 2 and 3 the sample of organizations from the EU countries and Serbia have been presented, which were involved in the CRANET research in the period from 2008 to 2010, and also the industry and sector distribution of the sample.

Table 2
The number of companies from EU and Serbia involved in the research

State	Frequency	Percent	Cumulative Percent
Austria	203	5.3	5.3
Belgium	240	6.2	11.5
Bulgaria	267	6.9	18.5
Cyprus	90	2.3	20.8
Czech Republic	54	1.4	22.2
Denmark	362	9.4	31.6
Estonia	74	1.9	33.6
Finland	136	3.5	37.1
France	157	4.1	41.2
Germany	420	10.9	52.1
Greece	214	5.6	57.7
Hungary	139	3.6	61.3
Ireland	103	2.7	64.0
Italy	157	4.1	68.0
Lithuania	119	3.1	71.1

Netherlands	116	3.0	74.1
Slovakia	225	5.9	80.0
Slovenia	219	5.7	85.7
Sweden	282	7.3	93.0
United Kingdom	218	5.7	98.7
Serbia	50	1.3	100.0
Total	**3845**	**100.0**	

Source: Authors' analysis

From Table 2 we can see the number of organizations that are included in the research.

Table 3
The industry in which companies from EU and Serbia operate

Industry	Frequency	%
Agriculture, hunting, forestry, fishing	82	2.1
Energy and water	138	3.6
Chemical products: extraction and processing of non-energy minerals	120	3.1
Metal manufacturing; mechanical, electrical and instrument engineering	509	13.2
Other manufacturing	482	12.5
Building and civil engineering	162	4.2
Retail and distribution; hotels; catering; repairs	334	8.7
Transport and communication	225	5.9
Banking; finance; insurance; business services	393	10.2
Personal, domestic, recreational services	31	.8
Health services	184	4.8
Other services	134	3.5
Education	145	3.8
Social services	63	1.6
Public administration	311	8.1
Other	386	10.0
Total	**3699**	**96.2**
Missing	146	3.8
Total	**3845**	**100.0**

Source: Authors' analysis

From Table 3 we can see the industry in which organizations from sample operate. The most presented industry is metal manufacturing, mechanical, electrical and instrument engineering (13%), other manufacturing (12.5%), banking, finance, insurance, and business services (10%) and retail, distribution, hotels and catering (8.7%).

4 Results of the Research

From Figure 1 we can see that outsourcing in the field of pension is used widely in EU countries, while in Serbia this external service is almost not used (M=0.15, SD=0.700). In Serbia compensation outsourcing is used mostly in the area of payroll (M=0.4, SD=1.127), while EU has even higher level usage of it (M=0.88, SD=1.429). Results from a nationwide survey identified payroll and employee benefits as two of the top five outsourced HR functions in Australia (Susomrith and Brown, 2013) which are similar with those explored in EU and Serbia. Generally, from the descriptive analysis technique we can conclude that all companies in each sample group use outsourcing for these compensation elements relatively modest (M min=0.08, M max=1.39 (out of 4.00)).

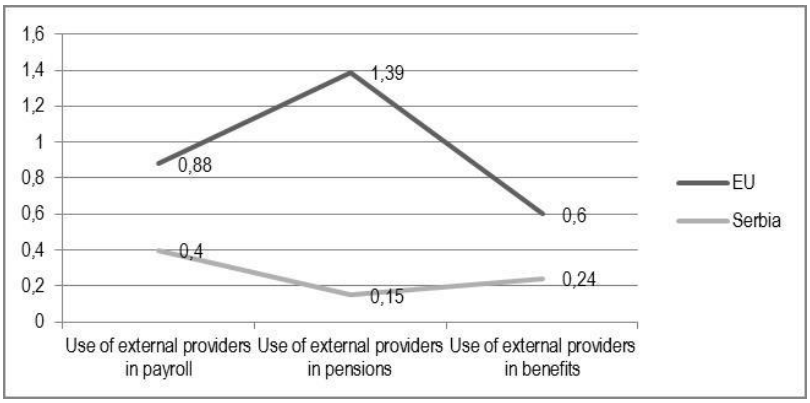


Figure 1
Comparative presentation of the usage of external providers for certain elements of compensation (statistical means) in EU and Serbia

It is important to mention one research from 2012 for the Central and Eastern Europe countries, where it was found that in Serbia external service providers were mostly used in training and development activities (70%). They were also often involved in recruitment (50%) and selection (55%), performance evaluation (50%) and at least in the area of compensation and benefits (45%). The practice of companies in this respect either have not changed a lot, or companies reported a decrease in use of these external partners [36], which is also in the line with given results. Serbian companies use outsourcing in a very small percentage, especially on pension (only 5%) and on benefits (7.3%). The largest usage of outsourcing is related to the payroll (14.3% of companies).

If we analyze the EU averages, we can see that there is a large gap between the EU and Serbia concerning their outsourcing practice. What may be the reason for it?

One answer may be that Serbia was the last country that entered the privatization process among other former Socialistic countries, with very high rate of unemployment and inflation, and relatively low level of economic growth. Besides, the social-political regime's changes after year 2000 influenced changes in the area of business, too, but there are still a lot of problems. A typical problem occurs in relation to the organization and management of the state-owned and public companies that are privatized (in very questionable manner, since the majority of those companies are unsuccessful today and they are existing on the edge of the bankruptcy) or are still state-owned and are going through restructuring process. Because of the underdeveloped market and general low level of professionalization in the area of human resource management in Serbia, HRO is used less than in the countries of EU. It has to be underlined that FDIs have many positive effects on Serbian economy (joint ventures, Greenfield, etc.). Foreign investors beside technology of production usually introduce totally new business concept especially in area of organization and management knowledge, and of course, in human resource management, too. One of these concepts is definitively HRO, which is used broadly in developed market economies.

In EU the average usage of payroll outsourcing is 32.1%; for benefits it is 29.8%; while for pension it is 48.7%. But there are also several differences among EU countries. For example, Belgium firms use outsourcing more than companies in any other EU country, so it will be very interesting to analyze these trends. In Figure 2 the graphical presentation of the usage of outsourcing of payroll, benefits and pension are given. It is obvious that the smallest divergence is in the area of benefit outsourcing, while the largest is in the area of pension administration. Many countries such as Belgium, Denmark, Finland, Austria, Finland, France, Germany, Italy, Sweden, Ireland, Netherland and UK use pension outsourcing between 40 and 80 percentage of companies. Then again, countries like Bulgaria, Czech Republic, Cyprus, Estonia, Greece and Lithuania use pension outsourcing only between 5 and 17 percentage of companies. This is a large diversity among EU countries. It has to be underlined that highly developed countries where are the headquarters of MNCs are, use outsourcing more than countries that are in the early stages of economic development (like Bulgaria, Czech Republic, Estonia, Slovenia, Slovakia, often called Central Eastern European countries). In the area of payroll and benefits there is smaller diversity in the usage of external providers.

Spearman's correlation was used to identify if there are any statistically significant relations between the level of outsourcing in the area of compensation and number of employees in company and number of employees in HR sector in those companies. According to the results of the correlation test for the EU sample - presented in table 4 - it is obvious that there is a strong positive correlation between payroll outsourcing and pension (r_s=0.443, p=0.000), payroll and benefits (r_s=0.408, p=0.000) and benefits and pension (r_s=0.425, p=0.000).

Table 4
Level of usage if external providers for HRM in EU (n=3795) and correlations

	Total number of employees	Total number employed by HR department	Use of external providers in payroll	Use of external provider in pension	Use of external providers in benefits
Total number of employees	(1.000)				
Total number employed by HR department	**.795****	(1.000)			
Use of external providers in payroll	**-.056****	.010	(1.000)		
Use of external providers in pensions	.016	.032	**.443****	(1.000)	
Use of external providers in benefits	.027	**.068****	**.408****	**.425****	(1.000)

**. Correlation is significant at the 0.01 level (2-tailed).

Source: Authors' analysis

The results of the non-parametric correlations (Spearman's rho) suggested that the relationship between outsourced elements of compensation, such as payroll, pension and benefits are statistically significant and positive - meaning that an increase in the usage of outsourcing of payroll indicates increase of outsourcing of pension and benefits. In the case of correlations analysis between the number of employees in the company and the existence HR department and the elements of compensation (payroll, benefits, pension) weak statistically significant negative correlations have been found between payroll and number of employees in companies (rs=-0.56, p=0.001) and weak positive correlation between number of employees in HR department (rs=0.068, p=0.00).

According to results of correlation test for Serbian sample – presented in Table 5 – it is obvious that a strong positive correlation exists between payroll outsourcing and pension (rs=0.334, p=0.035), payroll and benefits (rs=0.794, p=0.000) and benefits and pension (rs=0.500, p=0.001). The results of the non-parametric correlations (Spearman's rho) suggests that the relationship between outsourced elements of compensation such is payroll, pension and benefits are statistically

significant, positive and its means that an increase in the usage of outsourcing of payroll indicates an increase of outsourcing of pension and benefits. In the case of correlation analysis between the number of employees in the company and HR department and payroll, benefits and pension there were no statistically significant correlations found.

Table 5

Level of use if external providers for HRM in Serbia (n=50) and correlations

	Total number of employees	Total number employed by HR department	Use of external providers in payroll	Use of external provider in pension	Use of external providers in benefits
Total number of employees	(1.000)				
Total number employed by HR department	.212	(1.000)			
Use of external providers in payroll	-.020	.383	(1.000)		
Use of external providers in pensions	-.038	.021	**.334***	(1.000)	
Use of external providers in benefits	.074	.212	**.794****	**.500****	(1.000)

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

Source: Authors' analysis

To explore the influence of the number of employees, sector of business, industry, the existence of HR strategy and the usage of HRIS on HR outsourcing for payroll, benefits and pension, a multiple regression model was used.

In the analysis, a special attention was paid to the problems of multicollinearity, which is frequently present in the analyses due to their high inter-correlations.

SPSS achieved no multicollinearity (tolerance<.10 and VIF>10.0) in this model. In addition, there were no autocorrelation detected (Durbin-Watson coefficient was <2.00).

Table 6
Regression model for HR outsourcing in total sample

	Payroll			Pension			Benefits		
	B	**t**	**Sig.**	**B**	**t**	**Sig.**	**B**	**t**	**Sig.**
Const.	1,721	7,911	,000	,660	2,594	,010	,061	,360	,719
LN SIZE	-,066	-1,855	,064	,008	,204	,839	,000	-,007	,995
LN SIZE HRM	,037	,987	,324	,010	,231	,817	,047	1,630	,103
Sector	*-0.320*	*-4.304*	*0.000*	*0.198*	*2.303*	*0.021*	*-0.105*	*-1.817*	*0.069*
Industry	*-0.100*	*-1.738*	*0.082*	*-0.014*	*-0.205*	*0.838*	*-0.024*	*-0.544*	*0.586*
HRO pensions	*0.248*	*13.061*	*0.000*	*0.329*	*13.061*	*0.000*	*0.208*	*12.252*	*0.000*
HRO benefits	*0.349*	*12.252*	*0.000*	*0.478*	*14.804*	*0.000*	*0.215*	*14.804*	*0.000*
HR strategy	*0.089*	*2.328*	*0.020*	*-0.088*	*-1.995*	*0.046*	*0.051*	*1.728*	*0.084*
HRIS for pay	*-0.844*	*-8.418*	*0.000*	*-0.030*	*-0.258*	*0.797*	*0.064*	*0.808*	*0.419*
ANOVA	F	Sig	Df	F	Sig	Df	F	Sig	Df
	90.672	0.000	8	84.242	0.000	8	83.295	0.000	8
Model	R	R^2	Adj R^2	R	R^2	Adj R^2	R	R^2	Adj R^2
	0.524	0.275	0.272	0.51	0.260	0.257	0.508	0.258	0.255

Source: Authors' analysis

A multiple regression was run to predict HR outsourcing for payroll from number of employees in organization (log), number of employees in HR sector of organization (log), sector, industry, existence of HR strategy, the usage of HRIS for compensation, and the usage of HR outsourcing for other two elements of total compensation (pension and benefits). These variables statistically significantly predicted HR outsourcing for payroll, $F(8, 1916) = 90.672$, $p < 0.005$, $R2 = 0.275$. Variables: sector, HRO for pension, HRO for benefits, the existence of HR strategy and the usage of HRIS added statistically significantly to the prediction, $p < 0.05$.

In the case of HR outsourcing for pension the same variables statistically significantly predicted the usage of outsourcing, $F(8, 1916) = 84.242$, $p < 0.005$, $R2 = 0.260$. Variables: sector, HRO for payroll, HRO for benefits, the existence of HR strategy added statistically significantly to the prediction, $p < 0.05$.

A multiple regression was also run to predict HR outsourcing for benefits from number of employees in organization (log), number of employees in HR sector of organization (log), sector, industry, existence of HR strategy, the usage of HRIS for compensation, and the usage of HR outsourcing for other two elements of total compensation (pension and payroll). These variables statistically significantly

predicted HR outsourcing for payroll, $F(8, 1916) = 83.295$, $p < 0.005$, $R2 = 0.258$. Variables: HRO for pension and HRO for payroll only added statistically significantly to the prediction, $p < 0.05$.

Conclusions

In this paper the authors investigated, explained and provided examples and some current standards of human resource outsourcing (HRO), one of the key trends in the modern business. Implementing HRO practice, businesses may focus more on the core elements, by selling out the non-core elements, in order to reduce the cost and invest in the expertise of the core elements. Although outsourcing is described as beneficiary for organizations since it provides greater flexibility, which results in higher HR expertise, and better strategy with the elements essential to the business, from the analysis of samples from EU and Serbia, we can conclude that companies still do not use this possibility to a large extent which is in the line with other researches that have been done in the area of outsourcing and external providers in human resource management. The CRANET project provided a great sample of organizations suitable for this research.

Beside the fact that companies use HRO for payroll, pension and benefits at low level, it is important to emphasize that outsourcing is mostly used for payroll and pension as an administrative HR task, while benefits, which are today mostly interesting area in compensation in HRM, are outsourced in smaller percentage of companies in EU and Serbia. This is because benefits such flexible benefits, paternity leave, workplace child care, carrier break schemes, education break, cafeteria approach, etc. have great importance for employees and their motivation and satisfaction, so this can be a reason why many companies still do not use outsourcing for this special element of contemporary compensation package. Serbian companies use outsourcing at a very low level. This is influenced by slow and low level of economic development, in recent years and by insufficient knowledge and development in the area of HRM.

The results of the correlations suggested that the relationship between outsourced elements of compensation such is payroll; pension and benefits are statistically significant and positive, but weak or moderate (between 0 and 0.5; and 0.5 and 0.8). In the case of correlations analysis between the number of employees in the company and in HR department and payroll, benefits and pension there were no statistically significant correlations, so we did not confirmed these relations.

A multiple regression was run to predict HR outsourcing for payroll from number of employees in organization, number of employees in HR sector of organization, sector, industry, existence of HR strategy, the usage of HRIS for compensation, and the usage of HR outsourcing for other two elements of total compensation (pension and benefits). The variables of sector, HRO for pension, HRO for benefits, the existence of HR strategy and the usage of HRIS are statistically significant factors in the prediction of HR outsourcing for payroll. In case of HR outsourcing for pension the same variables statistically predicted the usage of

outsourcing, while HR outsourcing for benefits is predicted by HRO for pension and HRO for payroll.

Based on the theoretical background and presented empirical research, Human Resource Outsourcing may have significant benefits, but contemporary organizations interested in its implementation, should carefully analyze all data related to HRO, including all possible advantages and disadvantages. Since there are strong demands for cost cutting and downsizing, external providers are sometimes a very helpful solution. In contrast, the possible loss of loyalty, moral of employees and loss of knowledge-base can be quite a problem for future business development, which is predicted to be increasingly based on information, IT and knowledge.

References

[1] Alewell, D., Hauff, S., Thommes, K., & Weiland, K. (2009) Triggers of HR Outsourcing Decisions–an Empirical Analysis of German Firms. The International Journal of Human Resource Management, 20(7), 1599-1617

[2] Alfalla-Luque, R., García, J. A M., & Medina-López, C. (2012) Is Worker Commitment Necessary for Achieving Competitive Advantage and Customer Satisfaction when Companies Use HRM and TQM Practices?. Universia Business Review, (36), 64-89

[3] Baošić, M., Berber, N., Radičev, S., & Pasula, M. (2011) Human Resource Business Process Outsourcing: Trends and Challenges. In: Proceedings of the XV International Scientific Conference on Industrial Systems (IS'11), 14-16 September, Novi Sad: Faculty of Technical Sciences, 433-436

[4] Belcourt, M. (2006) Outsourcing—The Benefits and the Risks. Human Resource Management Review, 16(2), 269-279

[5] Berber, N., Pasula, M., Radosevic, M., Ikonov, D., & Kocic Vugdelija, V. (2012) Internal Audit of Compensations and Benefits: Tasks and Risks in Production Systems. Inzinerine Ekonomika-Engineering Economics, 23(4), 414-424

[6] Berber, N., Štangl Šušnjar, G., Slavić, A., & Baošić, M. (2014) Relationship between Corporate Social Responsibility and Human Resource Management-as New Management Concepts–in Central and Eastern Europe. Inzinerine Ekonomika-Engineering Economics, 25(3), 360-369

[7] Beregszaszi, J., & Polay, D. H. (2012) Human Resource Outsourcing in Times of Economic Turbulence–a Contemporary Review of Practice. International Journal of Human Resource Studies, 2(1), 46-65

[8] Björkman, I., & Smale, A. (2010) Global talent management: Challenges and Solutions. Universia Business Review, (27), 30-43

[9] Bonache, J., & Fernández, Z. (1997) Expatriate Compensation and Its Link to the Subsidiary Strategic Role: a Theoretical Analysis. International Journal of Human Resource Management, 8(4), 457-475

[10] Braun, I., Pull, K., Alewell, D., Störmer, S., & Thommes, K. (2011) HR Outsourcing and Service Quality: Theoretical Framework and Empirical Evidence. Personnel Review, 40(3), 364-382

[11] Brewster, C., Sparrow, P., & Vernon, G. (2007) International Human Resource Management. London: Chartered Institute of Personnel and Development

[12] Brewster, C., Mayrhofer, W., & Reichel, A. (2011) Riding the Tiger? Going along with Cranet for Two Decades—A Relational Perspective. Human Resource Management Review, 21(1), 5-15

[13] Briscoe, D. R., Schuler, R. S., & Claus, L. (2009) International Human Resource Management – Policies and Practice for Multinational Enterprises. London and New York: Routledge

[14] Buller, P. F., & McEvoy, G. M. (2012) Strategy, Human Resource Management and Performance: Sharpening Line of Sight. Human Resource Management Review, 22(1), 43-56

[15] Bustinza, O. F., Arias-Aranda, D., & Gutierrez-Gutierrez, L. (2010) Outsourcing, Competitive Capabilities and Performance: an Empirical Study in Service Firms. International Journal of Production Economics, 126(2), 276-288

[16] Chiang, F. F., Chow, I. H. S., & Birtch, T. A. (2010) Examining Human Resource Management Outsourcing in Hong Kong. The International Journal of Human Resource Management, 21(15), 2762-2777

[17] Cicek, I., & Ozer, B. (2011) The Effect of Outsourcing Human Resource on Organizational Performance: the Role of Organizational Culture. International Journal of Business and Management Studies, 3(2), 1131-1144

[18] Cooke, F. L., Shen, J., & McBride, A. (2005) Outsourcing HR as a Competitive Strategy? A Literature Review and an Assessment of Implications. Human Resource Management, 44(4), 413-432

[19] Delmotte, J., & Sels, L. (2008) HR Outsourcing: Threat or Opportunity? Personnel Review, 37(5), 543-563

[20] Dickmann, M., & Tyson, S. (2005) Outsourcing Payroll: Beyond Transaction-Cost Economics. Personnel Review, 34(4), 451-467

[21] Dowling, P. J., Festing, M., & Engle, A. D. (2008) International Human Resource Management – Managing People in a Multinational Context. London: Cengage Learning

[22] Dubravská, M., Mura, L., Kotulič, R., & Novotný, J. (2015) Internationalization of Entrepreneurship-Motivating Factors: Case Study of the Slovak Republic, Acta Polytechnica Hungarica, 12(5), 121-133

[23] Galanaki, E., & Papalexandris, N. (2007) Internationalization as a Determining Factor of HRM Outsourcing. The International Journal of Human Resource Management, 18(8), 1557-1567

[24] Gilley, K. M., Greer, C. R., & Rasheed, A. A. (2004) Human Resource Outsourcing and Organizational Performance in Manufacturing Firms. Journal of Business Research, 57(3), 232-240

[25] Hajdu, Z., Andrejkovič, M., & Mura, L. (2014) Utilizing Experiments Designed Results during Error Identification and Improvement of Business Processes, Acta Polytechnica Hungarica, 11(2), 149-166

[26] Huselid, M. A., & Becker, B. E. (2011) Bridging Micro and Macro Domains: Workforce Differentiation and Strategic Human Resource Management. Journal of Management, 37(2), 421-428

[27] Kakabadse, N., & Kakabadse, A. (2000) Critical Review–Outsourcing: a Paradigm Shift. Journal of Management Development, 19(8), 670-728

[28] Karoliny, Z., Farkas, F., & Poór, J. (2009) In focus, Hungarian and Central Eastern European Characteristics of Human Resource Management–An International Comparative survey. Journal for East European Management Studies, 14(1), 9-47

[29] Klaas, B. S. (2008) Outsourcing and the HR Function: an Examination of Trends and Developments within North American Firms. The International Journal of Human Resource Management, 19(8), 1500-1514

[30] Lee, R. P., & Kim, D. (2010) Implications of Service Processes Outsourcing on Firm Value, Industrial Marketing Management, 39(5), 853-861

[31] Lippert, S. K., & Swiercz, P. M. (2005) Human Resource Information Systems (HRIS) and Technology Trust, Journal of Information Science, 31(5), 340-353

[32] McIvor, R. (2008) What is the Right Outsourcing Strategy for your Process? European Management Journal, 26(1), 24-34

[33] Molnar, D., Vojtek, E., Borda, V., Szendro, K., & Juhasz, G. (2010) Evaluation Research on Outsourcing Human Resource Activities. Kaposvár: Human Exchange Human Resource Development and Consultant Foundation, 1-2, 91-110

[34] Papalexandris, N. (2005) Outsourcing of Human Resource Management Services in Greece. International Journal of Manpower, 26(4), 382-396

[35] Ployhart, R. E., & Moliterno, T. P. (2011) Emergence of the Human Capital Resource: A Multilevel Model. Academy of Management Review, 36(1), 127-150

[36] Poór, J., Nikolić, M., Slavić, A., & Štangl – Šušnjar, G. (2012) HRM under Changes at Foreign Subsidiaries in Serbia in Line With a Central and Eastern European Survey. Strategic Management, 17(1), 42-52

[37] Poór, J., Engle, A. D., Kovács, I. E., Slavić, A., Wood, G., Szabó, K., Stor, M., Kerekes, K., Karoliny, Z., Alas, R., & Némethy, K. (2015) HR Management at Subsidiaries of Multinational Companies in CEE in Light of Two Surveys of Empirical Research in 2008 and 2013. Acta Polytechnica Hungarica, 12(3), 229-249

[38] Raiborn, C. A., Butler, J. B., & Massoud, M. F. (2009) Outsourcing Support Functions: Identifying and Managing the Good, the Bad, and the Ugly. Business Horizons, 52(4), 347-356

[39] Ramayah, T. (2012) Determinants of Attitude towards E-HRM: an Empirical Study Among HR Professionals. Procedia-Social and Behavioral Sciences, 57, 312-319

[40] Seth, M., & Sethi, D. (2011) Human Resource Outsourcing: Analysis Based on Literature Review. International Journal of Innovation, Management and Technology, 2(2), 127-135

[41] Shih, H. A., & Chiang, Y. H. (2011) Exploring the Effectiveness of Outsourcing Recruiting and Training Activities, and the Prospector Strategy's Moderating Effect. The International Journal of Human Resource Management, 22(1), 163-180

[42] Štangl Šusnjar, G., & Leković, B. (2009) Performance-based Pay in Human Resources Development. Strategic Management, 14(3), 1-14

[43] Štangl Šušnjar, G., & Slavić, A. (2012) Changes in the Human Resource Compensation Systems of European Companies: Based on the CRANET research result analysis. Strategic Management, 17(4), 32-40, 2012

[44] Štangl Šušnjar, G., Slavić, A., & Berber, N. (2013a) The Analysis of Human Resource Outsourcing in Central and Eastern Europe. Metalurgia International, 18(11), 57-61

[45] Štangl-Šušnjar, G., Slavić, A., & Berber, N. (2013b) Human Resource Information Systems: Trends and Advantages. Metalurgia International, 18(8), 222-225

[46] Strohmeier, S. (2007) Research in e-HRM: Review and Implications. Human Resource Management Review, 17(1), 19-37

[47] Susomrith, P., & Brown, A. (2013) Motivations for HR Outsourcing in Australia. The International Journal of Human Resource Management, 24(4), 704-720

[48] Tremblay, M., Patry, M., & Lanoie, P. (2008) Human Resources Outsourcing in Canadian Organizations: An Empirical Analysis of the Role of Organizational Characteristics, Transaction Costs and Risks. The International Journal of Human Resource Management, 19(4), 683-715

[49] Turnbull, J. (2002) Inside Outsourcing. People Management: Connected HR, 10-11

[50] Van Buren III, H. J., Greenwood, M., & Sheehan, C. (2011) Strategic Human Resource Management and the Decline of Employee Focus. Human Resource Management Review, 21(3), 209-219

[51] Wright, P. M., & McMahan, G. C. (2011) Exploring Human Capital: Putting 'Human' Back into Strategic Human Resource Management. Human Resource Management Journal, 21(2), 93-104

[52] Yang, D. H., Kim, S., Nam, C., & Min, J. W. (2007) Developing a Decision Model for Business Process Outsourcing. Computers & Operations Research, 34(12), 3769-3778