computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

A NEW ALGORITHM FOR TIME SERIES DATA MINING BY USING ROUGH SET

Fei Hao [1], Ling Hei Yeung [2]

[1] Korea Advanced Institute of Science and Technology
373-1, guseong-Dong, Yuseong-Gu, Daejeon 305-701, Korea
fhao@islab.kaist.ac.kr
[2] Hong Kong Baptist University, Hong Kong
lightisgood2005@yahoo.com.hk

Abstract: *This paper is to apply Rough Set to data mining of time series. Firstly, we process the time series data by attribute selection and similarity sequence search. Secondly, the time series is partitioned into some sets of pattern by Mobile Window Method (MWM) and each pattern is a trend of time series. Thirdly, an information table is made by predicting attributes and targeting attribute in trending variation ratio structure sequence (TVRSS). Then, the original information table is made suitably for rough set to discover knowledge. Finally, the extracting rules can predict the time series behavior in the future. The total process is four steps. In the end, we show some examples to demonstrate our method on the time series data of stock market.*

Keywords: *TVRSS, Time series, Rough Set, Prediction.*

1. INTRODUCTION

Time-series data is becoming more and more vital in much field, especially in economic and finance area. With broad application of the information technology and diversification approach of obtaining data, time series information is increasing rapidly. People often try to mine the potential knowledge and useful information in some effective method or technologies when they confront a great deal of time series data. For example, consumers often make use of historical stock closing price data to predict future closing price. The procedure of converting historical data to useful knowledge or information for human is paid more attention in modern society.

Time-series data mining is an important way which mining some useful and potential knowledge from a great deal of time-series [1]. More and more people began to pay more attention to time-series data mining [2, 3, 4]. Nowadays, the study about time-series mining is mainly about similarity mining and time sequences query. Due to nonlinear, aperiodic and chaos of time series, it is difficult to predict future precision for human. But in a certain situation, it is possible to predict some potential information in time-series. For example, if you have necessary information about trend of stock closing price. We can predict the approximately trend in the future in spite of we can't predict the precise closing price.

In this paper, we mainly discussed the time-series data prediction based on TVRSS and rough set. Section 2 is devoted to introducing time-series and trending variation ratio structure sequence. Section 3 deals with some basic notions related to rough set theory (RST). In section 4, a novel time-series prediction method based on TVRSS and a time-series data mining model based on rough set are proposed respectively. Section 5 is experimental analysis. Conclusion is section 6.

2. TIME SERIES AND TVRSS

Time series is a series of observation data according to a certain time sequence [5]. It is a aggregate which has time and event. Time series is a data set in the planar or multidimensional space. Time-series data mining is a important way which mining some useful and potential knowledge from a great deal of time-series [6].

Time series possess characteristic as follows:

1. In a planar time series X, a certain point x_t is composed of abscissa time T and y-axis X (t, x) in planar space $T \times X$.
2. Time series is not reversible.
3. Information transfer is unilateral and not reversible; Information transfer direction is same as

the time. For example, a time series X={x_t|t=1, 2,…, n}. Information will transfer with $x_{t1}$to x_{t2} ($t_1<t_2$).

Definition 1 [5] *For a planar time series X={x_t|t=1, 2, …,n}, a certain point x_t is composed of abscissa time T and y-axis X (t, x) in planar space T × X. We can write time series in vector form*

$$X=\{x_1, x_2, \ldots, x_n\}$$

Definition 2 Let X={x_t|t=1, 2,…,n} be time series, λ={λ_1, λ_2,…, λ_{n-1}}.λ is a trending variation ratio structure sequence of X. In which, $\lambda i=\lambda i=\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}$, $\triangle^{i}_{x}=x_{i+1}-x_i$, $\triangle^{i}_{t}=t_{i+1}-t_i$ Obviously, λ is also a time series.

$$\lambda=\left\{\frac{\triangle^{1}_{x}}{\triangle^{1}_{T}},\frac{\triangle^{2}_{x}}{\triangle^{2}_{T}},\ldots,\frac{\triangle^{n-1}_{x}}{\triangle^{n-1}_{T}}\right\} \quad (1)$$

in which, I={1, 2,…, n-1}.

Definition 3 Let X = {x_t|t=1, 2, …, n} be time series, λ={λ_1, λ_2, …, λ_{n-1}} be trending variation ratio structure sequence of X. (x_i, x_{i+1},…, x_{i+k-1}), (λ_j, λ_{j+1},…, λ_{j+k+1}) (I=1, 2, …, n-k+1; j=1, 2, …, n-k) are called k=time sub series for X and λ respectively. {λ_h, λ_{h+1},…, λ_{h+k-1}} is a trending variation ratio structure sequence of time sub series {x_h, x_{h+1}, …, x_{h+k-1}} in λ. In which, $\lambda_m=\frac{\triangle^{m}_{x}}{\triangle^{m}_{T}}$, m=h, h+I,…, h+k-1 (h+k-1<n), xi∈X; {x_h, x_{h+1},…, x_{h+k-1}} is a latest time sub series which length is k of X.

k-time sub series is defined as follows,

$$S(X, k)=\{(x_i, x_{i+1},\ldots, x_{i+k-1})|(i=1,\ldots, n-k+1)\} \quad (2)$$

$$S(\lambda, k)=\{(\lambda_j, \lambda_{j+1},\ldots, \lambda_{j+k-1})|(j=1,2,\ldots, n-k)\} \quad (3)$$

Here, |S(X, *k*)|=n-k+1, S(λ, *k*)=n-k.

Definition 4 Suppose *s_1, s_2∈S(X, k)-{(x_{n-k+1}, x_{n-k+2},…, x_n)}, ι_1, ι_2∈S(λ,k), l_1,l_2 are trending variation ratio structure sequence of s_1,s_2 respectively. If d(ι_1, ι_2)=0, then the trending variation ratio structure of s_1 is same as s_2. In which, d(ι_1, ι_2) denotes the Euclid distance, d(ι_1, ι_2)=$\sum_i^k|\lambda_{1i}-\lambda_{2i}|$, ι_1={λ_{11}, λ_{12}, …, λ_{1k}}, ι_2={λ_{21}, λ_{22}, …, λ_{2k}}.*

Definition 5 *Suppose X={x_t|t=1, 2,…,n} be time series, $\triangle^{i}_{x}=x_{i+1}-x_i$, $\triangle^{i}_{T}=t_{i+1}-t_i$, i=1, 2, …, n-1, $\lambda=\left\{\frac{\triangle^{1}_{x}}{\triangle^{1}_{T}},\frac{\triangle^{2}_{x}}{\triangle^{2}_{T}},\ldots,\frac{\triangle^{n-1}_{x}}{\triangle^{n-1}_{T}}\right\}$ is the trending variation ratio structure sequence of X, then the trending variation ratio structure sequence of X are redefined as follows according to various $\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}$,*

$$\lambda^k=\begin{cases}\left(\left\{\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}|\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}>0\right\}\right)\\ \left(\left\{\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}|\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}=0\right\}\right)\\ \left(\left\{\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}|\frac{\triangle^{i}_{x}}{\triangle^{i}_{T}}<0\right\}\right)\end{cases} \quad i=1, 2, \ldots, n-1, k=1, 2, 3 \quad (4)$$

in which, $\triangle^{i}_{T}>0$ corresponding to ascending trending structure sequence λ^1. $\triangle^{i}_{T}=0$ corresponding to ascending trending structure sequence λ^2. $\triangle^{i}_{T}<0$ corresponding to ascending trending structure sequence λ^3.

3. DECISION-MAKING INFORMATION SYSTEM IN ROUGH SET THEORY

Information is a cognitive result about external world and a decision-making principle about human behavior. People can make some useful decision-making by historical knowledge, especially in prediction field and reasoning technology. Information system is a abstract description of database. Data in information system is usually denoted by relation table which row denotes objects and column denotes attributes. Information of objects is expressed by attribute value. It is an important tache between objects and attribute. It is also an information foundation for knowledge discovery.

Definition 6 [8] *Let U be a non-empty, finite set of objects called the universe, R be an equivalence relation on U. The pair (U, R) as approximation space. And a set of equivalent classes with respect to R is as follows*

$$U/R=\{[x_i]_R|x_i\in U\},$$

and where $[x_i]=\{x_j|(x_i, x_j)\in R\}$.

Definition 7 [8] *Information System (IS) is a quaternary presentation*

$$S=(U, A, V, f)$$

in which, U denotes a non-empty, finite set of objects called the universe, it is expressed as U={x_1, x_2, …, x_n}. A denotes a noe-empty, finite set of attribute it is expressed as A={a_1, a_2,…, a_m}. V is range of attribute. V={v_1, $_2$,…, v_m} where v_i is a value of a_i. f is an information function, f: U × A→V, $f(x_i, a_i)\in V_i$

Definition 8 [8] Decision Information System (DIS) is a quaternary presentation

$$S=(U, C\cup D, V, f)$$

where $A=C \cup D$ and $C\cap D\neq 0$ in which, C is the set of conditional attribute and D is the set of decision attribute.

Definition 9 [8] Each non-empty subset $p\subseteq A$ determines an indiscernibility relation as follows:

$$IND(p)=\{(x, y)\in U\times U|f_i(x)=f_i(y), a_i\in p\},$$

then IND(p) is an equivalence relation on U and let:

$$[x_i]_{IND(p)}=\{x_j|(x_i,x_j)\in IND(p)\}.$$

Property 1 IND(p) is an equivalence relation on U and $IND(p)=I_{a\in p}IND(a)$

Definition 10 Let S={U, A, V, f} be a decision-making information system, where $A=C\cup D$ and $C\cap D\neq 0$ $X_i=U/C$, $Y_j=U/D$ Decision-making rules are defined as follows:

$$r_{ij}: des(X_i)\rightarrow des(Y_j) \quad X_i\cap Y_j\neq 0 \qquad (5)$$

in which $des(X_i)$ denotes the description of each object on conditional attributes, and $des(Y_j)$ denotes the description of each object on decision-making attribute.

Corollary 1 Reliability gene of each rule is defined as follows,

$$\mu(X_i, Y_j)=|X_i\cap Y_j|/|X_i|$$

where, $0\leq\mu(X_i, Y_j)$

if $\mu(Xi,Yj)=1$, then r_{ij} is certain;

if $0<\mu(Xi,Yj)<1$, then r_{ij} is uncertain

Example 1 Give a decision-making table of insurance investigation in TaiWan (see Table 1), in which U={person1, person2,…, person8}={*x*1, *x*2,…, *x*8}, A={area, occupation, age, salary, insurance}={a1, a2, a3, a4, a5},

C={area, occupation, age, salary}={a1, a_2, a_3, a_4},

D={insurance}={a_5}

In above insurance investigation, conditional attributes and decision attribute are described as follows,

area={north, south}={1, 2};
Professional={worker, professor}={1, 2};
Age={young, middleage, old}={1, 2, 3};
Salary={low, high}={1, 2};
Insurance={living, poverty, house}={1, 2, 3};
Xi=U/C={{x1, x3}, {x2}, {x4, x5, x7}, {x6, x8}}
Yi=U/D={{x1, x2, x3}, {x4, x5, x7}, {x6, x8}}

For example, according to Eq. (4), we can obtain some useful decision-making rules easily:

Table 1. Insurance investigation

person	area	occupation	age	salary	insurance
x_1	1	1	3	2	1
x_2	2	1	2	1	1
x_3	1	1	3	2	1
x_4	1	2	2	1	2
x_5	1	2	2	1	2
x_6	1	2	1	1	3
x_7	1	2	2	1	2
x_8	1	2	1	1	3

Rule 1 If a person who lived in north of Taiwan, is a old worker, and have high salary, then he or she will buy living insurance;

Rule 2 If a person who lived in south of Taiwan and is middle age worker, **and** have low salary, then he or she will buy living insurance;

Rule 3 If a person who lived in north of Taiwan and is old worker, and have low salary, then he or she will buy living insurance;

Rule 4 If a person who lived in north of Taiwan and is middle age professor, and have high salary, then he or she will buy poverty insurance;

According to corollary 1, above rules are certain, because of their $\mu(X_i, Y_j)=1$.

4. TIME-SERIES DATA PREDICTION BASED ON TVRSS AND ROUGH SET

After introduced the basic concept of time series and decision-making information system in rough set theory. In this section, a model of time-series data mining based on rough set is established firstly, then a novel time-series data prediction method based on trending variation ratio structure sequence and rough set is proposed.

4.1. TIME-SERIES DATA MINING MODEL BASED ON ROUGH SET

Generally, data mining procedure is divided into steps as follows,

1. Establish data mining goal
2. Analysis and prepare data
3. Establish mining model according correlative algorithm
4. Evaluate mining model
5. Implement data mining

Without exception, time-series data mining model based on rough set contains data collection, data pretreatment, data reduction, extracting rules, classification and prediction [11].

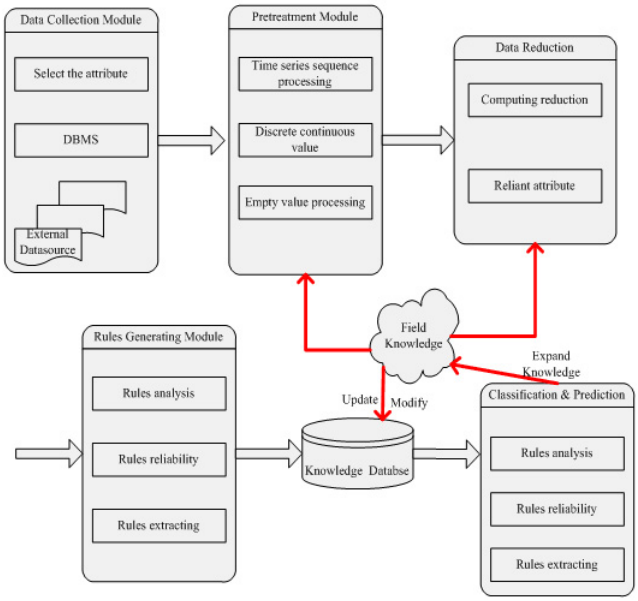


Fig. 1 – Time-series data mining model based on rough set

4.2. CONVERSION FROM TVRSS TO DECISION IN-FORMATION SYSTEM

Suppose time series $X=\{x_1,x_2,\ldots,x_n\}$, trending variation ratio structure sequence $\lambda=\{\lambda_1,\lambda_2,\ldots,\lambda_{n-1}\}$ Conversion step as follows:

step 1 Data standard pretreatment for λ by mobile window method [9].

step 2 Make trending variation ratio structure sequence convert into standard time-series data stylebook set.

step 3 Make standard time-series data stylebook set convert into decision-making information system.

Suppose mobile window size=k, we can construct trending variation ratio structure information system according to above steps, (see Table 2). Here, trending variation ratio structure information system reflects feature sub sequence property. So, we can analyze the time-series data with feature sub sequence.

Table 2. TVRSS information system

U\A	a_1	a_2	…	a_{k-2}	a_{k-1}	a_k
1	λ_1	λ_2	…	λ_{k-2}	λ_{k-1}	λ_k
2	λ_2	λ_3	…	λ_{k-1}	λ_k	λ_{k+1}
3	λ_3	λ_4	…	λ_k	λ_{k+1}	λ_{k+2}
4	λ_4	λ_5	…	λ_{k+1}	λ_{k+2}	λ_{k+3}
…	…	…	…	…	…	…
n-k	λ_{n-k}	λ_{n-k+1}	…	λ_{n-3}	λ_{n-2}	λ_{n-1}

Table 3. TVRSS predictable information system

U\A	a_1	a_2	…	a_{k-2}	a_{k-1}	a_k
1	λ_1	λ_2	…	λ_{k-2}	λ_{k-1}	λ_k
2	λ_2	λ_3	…	λ_{k-1}	λ_k	λ_{k+1}
3	λ_3	λ_4	…	λ_k	λ_{k+1}	λ_{k+2}
4	λ_4	λ_5	…	λ_{k+1}	λ_{k+2}	λ_{k+3}
…	…	…	…	…	…	…
n-k	λ_{n-k}	λ_{n-k+1}	…	λ_{n-3}	λ_{n-2}	λ_{n-1}
s_q	λ_{n-k+1}	λ_{n-k+2}	…	λ_{n-2}	λ_{n-1}	q

4.3. ONE-STEP PREDICTION BASED ON TVRSS PREDICTABLE INFORMATION SYSTEM

In time series $X=\{x_1,x_2,\ldots,x_n\}$, One-step prediction is that predict X_{n+1} value. But, it can't predict precise value, only predict the trend on time n+1. Time-series trending variation ratio structure in latest mobile window is $\{\lambda_{n-k+1}, \lambda_{n-k+2},\ldots,\lambda_{n-1}\}$, decision attribute value q is unknown, decision attribute value q is the prediction aim(q=-1,0,1) (see Table 3).

In Table3, different trending variation ratio depicts different pattern. These values are continuous, it is necessary to discrete these values according to its variation ratio. Here, fuzzy discrete method is adopted.

4.3.1 DATA PRETREATMENT

Definition 11 Suppose $X=\{x_t|t=1,2,\ldots,n\}$ be time series, the trending variation ratio structure $\lambda=(\frac{\Delta^1 X}{\Delta^1 T},\frac{\Delta^2 X}{\Delta^2 T},\ldots,\frac{\Delta^{n-1} X}{\Delta^{n-1} T})$, the membership functions of trending variation ratio on, $\lambda_1,\lambda_2,\lambda_3$ are defined respectively as follows,

$$\mu_{ascend_{rapid}}=\begin{cases}0, & |\lambda_i|\le b_1\\ \dfrac{|\lambda_i|-|b_1|}{a_1-b_1}, & b_1\le|\lambda_i|\le a_1\\ 1, & |\lambda_i|\ge a_1\end{cases}$$

$$\mu_{ascend_{moderate}}=\begin{cases}0, & |\lambda_i|\le b_2\\ \dfrac{|\lambda_i|-|b_2|}{a_2-b_2}, & b_2\le|\lambda_i|\le a_2\\ 1, & |\lambda_i|\ge a_1\end{cases}$$

$$\mu_{ascend_{slow}}=\begin{cases}0, & |\lambda_i|\le b_3\\ \dfrac{|\lambda_i|-|b_3|}{a_3-b_3}, & b_3\le|\lambda_i|\le a_3\\ 1, & |\lambda_i|\ge a_1\end{cases}\quad(6)$$

In Eq. (6), $\lambda_i\in\lambda^1$

$$\mu_{equal} = 1 \quad c \le |\lambda_i| \le d \qquad (7)$$

Here, $\lambda_i \in \lambda^2$

$$\mu_{descend_{rapid}} = \begin{cases} 0, & |\lambda_i| \le f_1 \\ \dfrac{|\lambda_i| - |f_1|}{e_1 - f_1}, & f_1 \le |\lambda_i| \le e_1 \\ 1, & |\lambda_i| \ge e_1 \end{cases}$$

$$\mu_{descend_{moderate}} = \begin{cases} 0, & |\lambda_i| \le f_2 \\ \dfrac{|\lambda_i| - |f_2|}{e_2 - f_2}, & f_2 \le |\lambda_i| \le e_2 \\ 1, & |\lambda_i| \ge e_2 \end{cases}$$

$$\mu_{descend_{slow}} = \begin{cases} 0, & |\lambda_i| \le f_3 \\ \dfrac{|\lambda_i| - |f_3|}{e_3 - f_3}, & f_3 \le |\lambda_i| \le e_3 \\ 1, & |\lambda_i| \ge e_3 \end{cases} \qquad (8)$$

In Eq. (8), $\lambda_i \in \lambda^3$

In which, $a_1,a_2,a_3,b_1,b_2,b_3,c,d,e_1,e_2,e_3,f_1,f_2,f_3$ is defined by users

The trending variation ratio structure sequence $л = \left\{ \frac{\Delta^1 X}{\Delta^1 T}, \frac{\Delta^2 X}{\Delta^2 T}, \ldots, \frac{\Delta^{n-1} X}{\Delta^{n-1} T} \right\}$ is redefined according to above membership functions of trending variation radio as follows,

$$\tilde{\lambda} = (\mu_1^1, \mu_2^1, ..., \mu_7^1, ..., \mu_1^{n-1}, \mu_2^{n-1}, ..., \mu_7^{n-1}) \qquad (9)$$

Moreover, Eq. (9) is formalized according to the maximum membership grade principle like this,

$$\text{e.g. } \tilde{\lambda} = (0,1,0,0,0,0,0|0,0,0,0,1,0,0| \\ 0,0,0,1,0,0,0|1,0,0,0,0,0,0) \qquad (10)$$

Eq. (10) depicts the feature of trending on variation ratio. So, its feature is explained as follows,

$$ascend_{moderate} \rightarrow descend_{rapid} \rightarrow equal \rightarrow ascend_{rapid}$$

Obviously, redefined $\tilde{\lambda}$ is constituted by seven-dimensional vector space. However, we can get the trend of future time according to the $\tilde{\lambda}$ corresponding to its state space vector.

4.3.2 PATTERN MATCHING

In this paper, we consider each trending structure sequence is a pattern.Due to this idea, Hamming distance is adopted in this paper.

Definition 12 *Suppose a certain pattern* $l_{n-k} = (\lambda_{n-k}, \lambda_{n-k+1}, ..., \lambda_{n-1}) \in S(\lambda, k)$, *the trending structure state space vector of* l_{n-k} *is denoted as follows,*

$$\tilde{\lambda}^{n-k} = (p_{n-k1}, p_{n-k2}, ..., p_{n-kn-k+1}) \\ k=1,2...,n-k+1 \qquad (11)$$

Definition 13 Let $s_1, s_2 \in S(X,k) - \{(x_{n-k+1}, ..., x_n)\}, l_1, l_2, \in S(\lambda, k), l_1, l_2$ are the time sub series of $s_1 m, s_2$ respectively. $\tilde{\lambda}^{(1)}, \tilde{\lambda}^{(2)}$ are the trending structure state space vector of l_1, l_2. Hamming distance between $\tilde{\lambda}^{(2)}$ and $\tilde{\lambda}^{(2)}$ is defined as follows,

$$H(\tilde{\lambda}^{(1)}\tilde{\lambda}^{(2)} = \sum_{i=1}^{n-k+1} h_i$$

$$h_i = \begin{cases} 1, & p_i \ne p_2, \lambda_{1i}\lambda_{2i} > 0 \\ 0, & \tilde{\lambda}^{(1)} = \tilde{\lambda}^{(2)} \\ 2, & p_i \ne p_2, \lambda_{1i}\lambda_{2i} < 0. \end{cases} \qquad (12)$$

in which, $l_1 = (\lambda_{11}, ..., \lambda_{1k}), l_2 = (\lambda_{21}, ..., \lambda_{2k})$

Property 2 $0 \le H(\tilde{\lambda}^{(1)}, \tilde{\lambda}^{(2)}) \le 2k$

If $H(\tilde{\lambda}^{(1)}, \tilde{\lambda}^{(2)}) = 0$, then K-time sub sequence s_1 has the same trending variation ratio with s_2.

The smaller $H(\tilde{\lambda}^{(1)}, \tilde{\lambda}^{(2)})$ is, the better pattern matching is.

Here, we give a pattern difference measurement method, we can easily get the same or similar pattern with goal pattern (predictable sub time sequence). In a intuitionistic sense, $H(\tilde{\lambda}^{(1)}, \tilde{\lambda}^{(2)})$ reflects the synthetical similarity between l_1 and l_2 $i = 1, 2..., n-k$. So, obtaining the relationship and extracting useful rules among these patterns is a problem.

The pattern matching model is devised as follows,

$$Min\ H(s_q, \tilde{\lambda}^{(i)})$$

$$s.t.\ \lambda_{si}\lambda_{ji} > 0$$

$$i = 1,2,...n-k$$

4.3.3 EXTRACTING PREDICTABLE RULES

In fact, each object in Table3 denotes a known predictable rule. These predictable rules are wrote like this according to Definition 10.

$$eg.if\ \text{descend}_{\text{rapid}} \rightarrow ascend_{moderate} \\ \rightarrow ascend_{rapid} \rightarrow ascend_{slow} \rightarrow ascend_{moderate} \\ \text{the } decend_{slow}$$

In the trending variation ratio structure predictable information system (See table 3), q is our prediction aim, s_q is predictable pattern, we can

predict the q by conditional expression C={$\lambda_{n-k+1}, \lambda_{n-k+2}, \ldots, \lambda_{n-1}$} in s_q.

Definition 14 *Let Sim(x, y) be the function of the similarity about x and y. For prediction sequence s_q and l_i, its similarity is defined as follows*

$$Sim(s_q, l_i) = \frac{1}{H(s_q, l_i)} \tag{13}$$

in which q=-1,0,1;i=1,2,…,n-k.

Note 1 The smaller $H(s_q, l_i)$ is, the better $Sim(s_q, l_i)$ is, i.e, the trusty q is.

Remark: The feature of the algorithm of trending variation ratio structure sequence based on rough set is concluded as follows,

Consider the conditional attribute sets belong to in-discernibility equivalence relation on objects, but not other objects. Hence, calculated capacity is lesser.

Description of trend structure classification is subtle.

Fuzzy discrete method accords with human intuitivism.

4.4 ALGORITHMS

A: Algorithm of Data Conversion and Resolution on $S(\lambda, k)$

Step1: Input $X = \{x_1, x_2, ...x_n\}$

Step2: Computing TVRSS $\lambda = (\frac{\Delta_X^1}{\Delta_T^1}, \frac{\Delta_X^2}{\Delta_T^2}, ..., \frac{\Delta_X^{n-1}}{\Delta_T^{n-1}}) = (\lambda_1, \lambda_2, ..., \lambda_{n-1})$ according to $\lambda_i = \frac{\Delta_X^i}{\Delta_T^i}$.

Step3: $S(\lambda, k) \leftarrow \phi$

For (t=1; t<n-k; t++) {

$U(t) = (\lambda_1, \lambda_2, ...\lambda_{t+k-1})$;

$S(T_r, k) \leftarrow S(T_r, k) \cup \{U(t)\}$

}

Step4: output $S(\lambda, k)$

B: Adjacence Algorithm of Computing $S(\lambda, k-1)$ and l_{n-1}

Step1: Input $S(\lambda, k-1)$

$M \leftarrow 0$

Step2. For ($i = 1; i < |S(\lambda, k-1)|; i++$) {

$\tilde{\lambda} \leftarrow F(\lambda)$

For $\lambda^{n-k} \in \tilde{\lambda}$, calculating $H(s_q, \lambda^{n-k})$

$Min(H(s_q, \lambda^{n-k}))$

}

5. EXPERIMENT

As we all known, stock data is a typical time series database. Stock code, opening price and closing price for every day are kept in the stock database. Here, stock code is static attribute, but opening price and closing price are dynamic attributes which change with time. Investor pay more attention to the alteration of opening price and closing price. In this paper, closing price is considered and analyzed.

We make short-term closing price forecast of Shen zhen Developing Bank (China) stock data according to the 32 transaction data (from March 1, 2005 to April 13, 2005, see Table4) [10].

5.1 EXPERIMENT STEP

Step 1: After computing trending variation ratio structure sequence, we can compute the trending variation ratio structure sequence of stock time series according to definition 2. (see table 4)

The stock series is expressed as follows according to definition 1:

X={6.48, 6.40, 6.38, 6.44, 6.33, 6.38, 6.36, 6.28, 6.10, 6.12, 6.10, 5.99, 5.94, 6.0, 5.93, 5.99, 5.54, 5.37, 5.40, 5.38, 5.30, 5.30, 5.09, 5.20, 5.52, 5.42, 5.5, 5.7, 6.29, 6.8, 6.88, 6.7}

Table 4. Transaction data (t, p denote time and closing price respectively)

t	1	2	3	4	5	6	7	8
p	6.48	6.40	6.38	6.44	6.33	6.38	6.36	6.28
t	9	10	11	12	13	14	15	16
p	6.10	6.12	6.10	5.99	5.94	6.0	5.93	5.99
t	17	18	19	20	21	22	23	24
p	5.54	5.37	5.40	5.38	5.30	5.30	5.09	5.20
t	25	26	27	28	29	30	31	32
P	5.52	5.42	5.5	5.7	6.29	6.8	6.88	6.7

Trending variation ratio structure sequence of stock time series is,

λ ={-0.02, 0.06, -0.08, -0.04, -0.11, -0.06, -0.13, -0.04, -0.0133, -0.09, -0.06, 0.04, -0.04, 0.0033, -0.4, -0.18, 0.04, 0, -0.0333, -0.2, 0.11, 0.39, -0.06, 0.07, 0.12, 0.56, 0.62, 0.0967, -0.44, 0.24}

Step 2: Let mobile windows size k=5, k-time sub series are obtained according to definition 3. (see Table 5)

Step 3: Convert TVRSS into predictable decision-making information system according to table 3 (see Table 5). Here, q is our prediction aim.

Step 4: Fuzzy discrete trending variation ratio according to Data pretreatment method in section.4.3.1. (see table 6)

Step 5: Computing the Hamming distance between s_q and l, i.e, $H(s_q, l)$. Due to definition 13, $H(s_q, l)$ are solved as follows (see Table 7).

Step 6: Computing the maximal similarity degree of Hamming distance according to Eq. (13).

$$H(s_q, s_{25}) = \frac{1}{3}$$

Finally, we learnt similarity degree of Hamming distance according is maximal when $s = s_{25}$. So, the similarity is the best among all feature sub sequences. Hence, the rules of sequence s_{25} is inserted in knowledge database, helping make future prediction.

Because of above analysis, we can get the experimental result that the closing price at time 33 is higher than its former closing price. In fact, observing closing price at time 33 is higher than its former closing price. From the decision-making rule point of view, we can gain such rule as follow,

Decision-making rule If $A_1 = ascend_{\mathrm{mod}erate}, A_2 = ascend_{apid}, A_3 = descend_{\mathrm{mod}erate}, A_4 = ascend_{\mathrm{mod}erate}$, then $D = ascend_{\mathrm{mod}erate}$. i.e, If closing price of the stock market keeps rising for five days and declines next day, then the closing price will be higher than former closing price in the future.

Table 5. Trending prediction table of stock time series

	A_1	A_2	A_3	A_4	D
s_1	-0.02	0.06	-0.08	-0.04	0.11
s_2	0.06	-0.08	-0.04	0.11	-0.06
s_3	-0.08	-0.04	0.11	-0.06	-0.13
s_4	-0.04	0.11	-0.06	-0.13	-0.04
s_5	0.11	-0.06	-0.13	-0.04	-0.0133
s_6	-0.06	-0.13	-0.04	-0.0133	-0.09
s_7	-0.13	-0.04	-0.0133	-0.09	-0.06
s_8	-0.04	-0.0133	-0.09	-0.06	0.04
s_9	-0.0133	-0.09	-0.06	0.04	-0.04
s_{10}	-0.09	-0.06	0.04	-0.04	0.0033
s_{11}	-0.06	0.04	-0.04	0.0033	-0.4
s_{12}	0.04	-0.04	0.0033	-0.4	-0.18
s_{13}	-0.04	0.0033	-0.4	-0.18	0.04
s_{14}	0.0033	-0.4	-0.18	0.04	0
s_{15}	-0.4	-0.18	0.04	0	-0.0333
s_{16}	-0.18	0.04	0	-0.0333	0
s_{17}	0.04	0	-0.0033	0	-0.2
s_{18}	0	-0.0033	0	-0.2	0.11
s_{19}	-0.0033	0	-0.2	0.11	0.39
s_{20}	0	-0.2	0.11	0.39	-0.06
s_{21}	-0.2	0.11	0.39	-0.06	0.07
s_{22}	0.11	0.39	-0.06	0.07	0.12
s_{23}	0.39	-0.06	0.07	0.12	0.56
s_{24}	-0.06	0.07	0.12	0.56	0.62
s_{25}	0.07	0.12	0.56	0.62	0.0967
s_{26}	0.12	0.56	0.62	0.0967	-0.44
s_{27}	0.56	0.62	0.0967	-0.44	0.24
s_q	0.62	0.0967	-0.44	0.24	goal

Table 6. Trending prediction table of stock time series

	A_1	A_2	A_3	A_4	D
s_1	0000001	0010000	0000010	0000001	0100000
s_2	0010000	0000010	0000001	0100000	0000010
s_3	0000010	0000001	0100000	0000010	0000010
s_4	0000001	0100000	0000010	0000010	0000001
s_5	0100000	0000010	0000010	0000001	0000001
s_6	0000010	0000010	0000001	0000001	0000010
s_7	0000010	0000001	0000001	0000010	0000010
s_8	0000001	0000001	0000010	0000010	0010000
s_9	0000001	0000010	0000010	0010000	0000001
s_{10}	0000010	0000010	0010000	0000001	0010000
s_{11}	0000010	0010000	0000001	0010000	0000100
s_{12}	0010000	0000001	0010000	0000100	0000100
s_{13}	0000001	0010000	0000100	0000100	0010000
s_{14}	0010000	0000100	0000100	0010000	0000100
s_{15}	0000100	0000100	0010000	0000100	0000001
s_{16}	0000100	0010000	0000100	0000001	0001000
s_{17}	0010000	0000100	0000001	0001000	0000100
s_{18}	0000100	0000001	0001000	0000100	0100000
s_{19}	0000001	0001000	0000100	0100000	1000000
s_{20}	0001000	0000100	0100000	1000000	0000010
s_{21}	0000100	0100000	1000000	0000010	0100000
s_{22}	0100000	1000000	0000010	0100000	0100000
s_{23}	1000000	0000010	0100000	0100000	1000000
s_{24}	0000010	0100000	0100000	1000000	1000000
s_{25}	0100000	0100000	1000000	1000000	0100000
s_{26}	0100000	1000000	1000000	0100000	0000100
s_{27}	1000000	1000000	0100000	0000100	1000000
s_q	1000000	0100000	0000100	1000000	goal

Table 7. H(sq, si) table

s	$H(s_q, s_i)$	s	$H(s_q, s_i)$	s	$H(s_q, s_i)$
s1	2+1+1+2=6	s_2	1+2+1+1=5	s_3	2+2+2+2=8
s4	2+0+1+2=5	s_5	1+2+1+2=6	s_6	2+2+1+2=7
s7	2+2+1+2=7	s_8	2+2+1+2=7	s_9	2+2+1+1=6
s10	2+2+2+2=8	s_{11}	2+1+1+1=5	s_{12}	1+2+2+2=7
s13	2+1+0+2=5	s_{14}	1+2+0+1=4	s_{15}	2+2+2+2=8
s16	2+1+2+2=7	s_{17}	1+2+1+2=6	s_{18}	2+2+2+2=8
s19	2+2+0+1=5	s_{20}	2+2+2+0=6	s_{21}	2+0+2+2=6
s22	1+1+1+1=4	s_{23}	0+2+2+1=5	s_{24}	2+0+2+0=4
s25	1+0+2+0=3	s_{26}	1+1+2+1=5	s_{27}	0+1+2+2=5

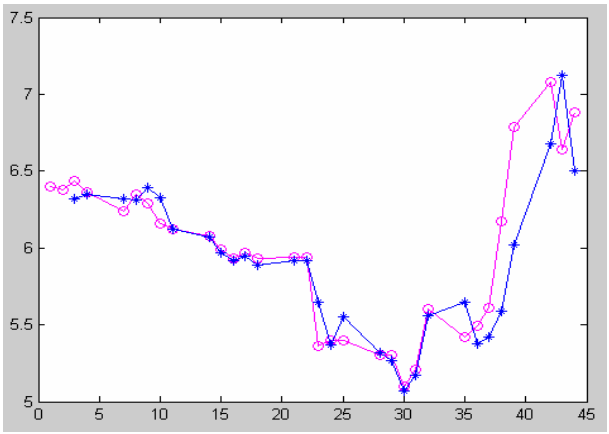
5.2 EXPERIMENT RESULT

In this paper, we utilized different Mobile window size to calculating the prediction result and accuracy.

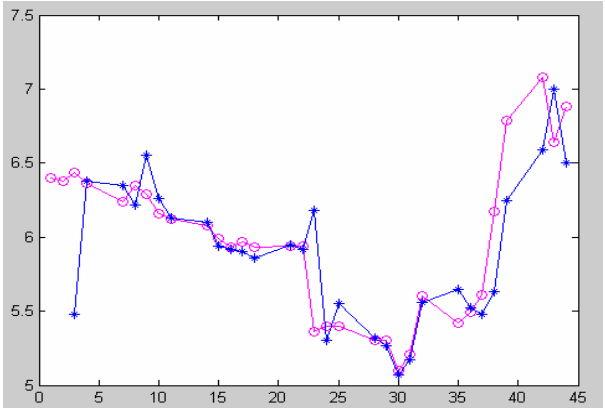
Table 8. Pattern results for different mobile window sizes

Factors / Size	Hamming Distance	Matching No	Prediction Results	Confidence
$k=3$	0+1=1	2	Descend moderate	50%
	0+1=1		Descend slow	50%
$k=4$	1+1+1=3	2	Ascend moderate	50%
	1+1+1=3		Ascend rapid	50%
$k=5$	1+1+1+1 =4	1	Ascend moderate	100%
$k=6$	#	0	#	
$k=7$	#	0	#	

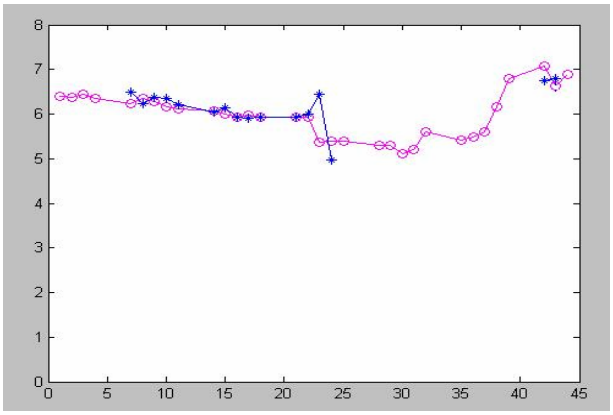
So, we can easily get the optimal mobile window size=3, and the prediction results (see Fig2.(a)).



(a) Size=3



(b) Size=4



(c) Size=5

Fig. 2 – Actual curve and prediction curve for different mobile window sizes

6. CONCLUSION

This paper mainly discussed the time-series data prediction based on trending variation ratio structure sequence and rough set. To obtain various k-time series, mobile window method is adopted. As a matter of fact, each sub time series is pattern. Trending variation ratio structure sequence is applied to sub time series in order to get important attribute which depict each sub time series. To predict unknown data or information in the future, trending variation ratio structure predictable information system is constructed. Hamming distance is adopted in pattern identification. Our method is applied to stock market data prediction.

7. REFERENCES

[1] Shao Fengjing, Yu Zhongqing. *Principle and Algorithm of Data Mining,* Water conservancy & water electric press of China, Beijing, 2003.

[2] P.Lee J, S.Kim. "Trend Similarity and Prediction in Time-series Databases [A]". *In: Proc. of SPIE on Data Mining and Knowledge Discovery: Theory, Tools, and Technology II. Washington: SPIE,* 2000,pp.201-212

[3] R.Agrawal, C.Faloutsos, A.Swami. "Efficient Similarity Search in Sequence Database". *Springer Verlag,* 1993, pp.69-84.

[4] G.Das, D.Gunopulous, H.Mannila. "Finding Similar Time Series", *In: Proc. of 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD97),* Komorowski J, Zytkow J (Eds.), 1997.

[5] Z.Stefan. Data Mining for Prediction. Financial Series Case. *Doctoral thesis, Department of Computer and System Sciences.* The Royal Institute of Technology, 2003.

[6] R.J.Bayardo Jr., R.Agrawal. "Mining the Most Interesting Rules", *In: Proc. of the 5th ACM*

SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining, August 1999.

[7] Yong Wang, Time-series data mining study and its application in prediction of water quality. *Doctor degree dissertation, Department of automatization.* Guangdong university of technology, 2005.

[8] Z. Pawlak, "Rough sets", *International Journal of Computer and Information Sciences,* vol.11, 1982, pp.341-356.

[9] J.K.Baltzersen. An attempt to predict stock data: a rough sets approach. *Diploma thesis, Konwledge Systems Group, Department of Computer Systems and Telematics,* The Norwegian Institute of Technology, University of Trondheim, 1996.

[10] http://quote.stock.163.com/h/data/dayhtml/000001.htm

[11] Keyun Hu, *Research and design of knowledge discovery system based on rough set,* HeFei university of technology, 1998.

Fei Hao *received the bachelor's degree & Master's degree in school of Mathematics & Computer Engineering from Xihua university in 2005 and 2008 respectively. Currently he is a PHD candidate in department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea.*

His research interests include intelligent information processing, rough sets theory and time series data mining, data stream.

Ling Hei Yeung *graduated 2004 ~ 2007 Bsc (Hons) in Computing Mathematics, Department of Mathematics, City University of Hong Kong. 2007~ 2009 Double Msc in Operational Research and Business Statistics, Jointly by Hong Kong Baptist University and Kent University of UK.*

Distinction: Dean's List, 2005, Fall Semester, City University of Hong Kong.

Research Interests: Operational Research and Business Statistics, Data Mining.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

AN ITERATIVE METHOD FOR THE EVALUATION OF THE REGULARIZATION PARAMETER IN REGULARIZED IMAGE RESTORATION

Said E. El-Khamy [1], Mohiy M. Hadhoud [3], Moawad I. Dessouky [2], Bassiouny M. Salam [2], Fathi E. Abd El-Samie [2]

[1] Department of Electrical Engineering, Faculty of Engineering,
Alexandria University, Alexandria, 21544, Egypt
[2] Department of Electronics and Electrical Communications, Faculty of Electronic Engineering,
Menoufia University, 32952, Menouf, Egypt
[3] Department of Information Technology, Faculty of Computers and Information,
Menoufia University, 32511, Shebin Elkom, Egypt
E-mails: elkhamy@ieee.org, mmhadhoud@yahoo.com, dr_moawad@yahoo.com,
b_m_salam@yahoo.com and fathi_sayed@yahoo.com

Abstract: *Regularized restoration is one of the powerful image restoration techniques because it preserves image details with a high degree of fidelity in the restored image. The main problem encountered in regularized image restoration is the evaluation of the regularization parameter. There are several methods for the evaluation of this parameter which require knowledge of the noise variance in the degraded image. After evaluating this parameter, regularized restoration is implemented by applying a regularization filter on the degraded image. In this paper, we propose a new iterative method for the evaluation of this parameter. This method depends on the maximization of the power in the restored image by the coincidence of the passband of the regularization filter with the frequency band in which, most of the image power exists. The suggested method doesn't require a priori knowledge of the noise variance. Results show that the estimated value of the regularization parameter leads to a minimum mean square restoration error.*

Keywords: *regularized restoration, regularization parameter, iterative regularization, inverse regularization.*

1. INTRODUCTION

Digital image restoration is a problem which has been extensively treated in the literature [1-12]. The main objective of image restoration is to obtain a good estimate of the original image from a degraded image. Degradations in images have several origins such as out of focus blurring, linear motion blurring and Gaussian blurring. These types of blurring can be modeled as lowpass filters affecting the original image [13]. So, the image restoration problem is in fact a deconvolution problem. The existence of noise in the degraded image increases the difficulty of the image restoration process. The problem of deconvolution in the presence of noise is classified as an ill-posed inverse problem [6].

This problem has been solved in the literature using so many approaches [1-13]. In each approach, the mathematical basis on which the solution is based differs. In general, the purpose of image restoration is to obtain an estimate, which is as close as possible to the original image. One of the most popular approaches to the problem of image restoration is the linear minimum mean square error (LMMSE) restoration approach [13]. Rather than seeking a solution consistent with minimum contamination by noise, the LMMSE approach attacks the restoration problem directly and proposes a criterion that explicitly evaluates how close the restoration is to the original object intensity distribution. Despite being an easy solution, LMMSE restoration requires some assumptions related to the original unknown image and leads to some artifacts in flat areas of the restored images.

Regularization theory, which was basically introduced by Tikhonov and Miller has proved to be a good candidate for the solution of the image restoration problem [5]. There is no guarantee for the existence, uniqueness and stability of the solution of an inverse ill-posed problem like the image restoration problem based on direct inversion. So, there is a need for some constraints on the

solution. The stabilizing functional approach is one of the basic methodologies for the development of regularized or constrained solutions. According to this approach, an ill-posed problem can be formulated as the constrained minimization of a certain functional, called the stabilizing functional [5].

For the implementation of regularized image restoration, there is a need to know both the regularization operator and the regularization parameter [1-12]. The rule of the regularization operator is to move the small Eigenvalues of the image degradation matrix away from zero while leaving the large Eigenvalues unchanged. It also incorporates prior knowledge about the required degree of smoothness of the restored image into the restoration process. The regularization parameter controls the trade-off between fidelity of the data and the smoothness of the solution [5, 6]. Hence, its determination is a very important issue. The evaluation of the regularization parameter can be performed using several techniques but most of these techniques are based on the amount of noise in the degraded image [6].

In this paper, we propose an iterative method for the evaluation of the regularization parameter based on the maximization of the total power content of the restored image. The rest of the paper is organized as follows. Section 2 gives the image degradation model. Section 3 discusses regularized image restoration. Section 4 surveys some existing methods for the evaluation of the regularization parameter. Section 5 presents the proposed method for the evaluation of the regularization parameter. Section 6 gives the experimental results. Finally, the concluding remarks are given in section 7.

2. IMAGE DEGRADATION MODEL

Image restoration algorithms are, generally, designed to exploit characteristics of an image and its degradation. Accurate knowledge of the degradation is essential for any image restoration algorithm to be a successful algorithm. To obtain information about the degradation, we can gather information from the degraded image itself. As an example, if an image is blurred and we can identify a region in the degraded image where the original undegraded signal is known, we may be able to estimate the blurring function $h(n_1, n_2)$. The original image is denoted by $f(n_1, n_2)$, while the degraded image is denoted by $g(n_1, n_2)$. In the presence of noise $n(n_1, n_2)$, this can be represented as follows [1-13]:

$$g(n_1, n_2) = h(n_1, n_2) * f(n_1, n_2) + n(n_1, n_2) \qquad (1)$$

If $f(n_1, n_2)$ is assumed to be known in some regions, then $g(n_1, n_2)$ and $f(n_1, n_2)$ can be used in these regions to estimate $h(n_1, n_2)$. This linear shift invariant image degradation model can be represented in another form as follows [1-13]:

$$\mathbf{g} = \mathbf{H}\mathbf{f} + \mathbf{n} \qquad (2)$$

where $\mathbf{f}$, $\mathbf{g}$ and $\mathbf{n}$ are lexicographic orders by either column or row of the M x M image, the degraded image and the noise, respectively. The matrix $\mathbf{H}$ is the discrete representation of the degradation of dimensions M^2 x M^2. For linear shift invariant systems, the matrix $\mathbf{H}$ is a block Toeplitz matrix. The objective of image restoration is to estimate $\mathbf{f}$ given the samples of the recorded image $\mathbf{g}$.

3. REGULARIZED IMAGE RESTORATION

According to the regularization theory, the solution of Eq. (2) is obtained by the minimization of the cost function [1-12]:

$$\Psi(\hat{\mathbf{f}}) = \left\| \mathbf{g} - \mathbf{H}\hat{\mathbf{f}} \right\|^2 + \lambda \left\| \mathbf{C}\hat{\mathbf{f}} \right\|^2 \qquad (3)$$

where $\mathbf{C}$ is the regularization operator and λ is the regularization parameter. This minimization is accomplished by taking the derivative of the cost function yielding [1-12]:

$$\frac{\partial \Psi(\hat{\mathbf{f}})}{\partial \hat{\mathbf{f}}} = \mathbf{0} = 2\mathbf{H}^t(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}) - 2\lambda \mathbf{C}^t\mathbf{C}\hat{\mathbf{f}} \qquad (4)$$

Solving for that $\hat{\mathbf{f}}$ that provides the minimum of the cost function gives [1-12]:

$$\hat{\mathbf{f}} = (\mathbf{H}^t\mathbf{H} + \lambda \mathbf{C}^t\mathbf{C})^{-1}\mathbf{H}^t\mathbf{g} = \mathbf{A}(\lambda)\mathbf{g} \qquad (5)$$

where

$$\mathbf{A}(\lambda) = (\mathbf{H}^t\mathbf{H} + \lambda \mathbf{C}^t\mathbf{C})^{-1}\mathbf{H}^t \qquad (6)$$

The regularization operator $\mathbf{C}$ incorporates prior knowledge about the required degree of smoothness of $\mathbf{f}$ into the restoration process. It can be a finite difference matrix chosen to minimize the second order or higher order difference energy of the estimated image [1]. The 2-D Laplacian is the best choice for this purpose. The rule of the regularization operator $\mathbf{C}$ is to move the small Eigenvalues of $\mathbf{H}$ away from zero while leaving the large Eigenvalues unchanged. The amount of change of these Eigenvalues is dependent on the choice of the regularization parameter λ.

Equation (5) can be written in an equivalent form using the Toeplitz to circulant approximation as follows [1]:

$$\hat{F}(u,v) = \frac{H^*(u,v)}{|H(u,v)|^2 + \lambda|C(u,v)|^2} G(u,v) = D(u,v,\lambda)G(u,v) \quad (7)$$

where $F(u,v)$ is the Fourier transform of the original image, and $G(u,v)$, $\hat{F}(u,v)$, $H(u,v)$, $C(u,v)$ and $N(u,v)$ are the Fourier transforms of the degraded image, the estimate of the original image, the point spread function PSF of the blurring operator, the regularization operator and the noise, respectively. Thus, the transfer function of the regularization filter is given by [1]:

$$D(u,v,\lambda) = \frac{H^*(u,v)}{|H(u,v)|^2 + \lambda|C(u,v)|^2} \quad (8)$$

If λ=0, this leads to the inverse filter solution defied as [1]:

$$\hat{F}(u,v) = \frac{G(u,v)}{H(u,v)} = F(u,v) + \frac{N(u,v)}{H(u,v)} \quad (9)$$

In the presence of noise, the restoration problem has an ill-posed nature. Thus, severe deteriorations are observed in the restored images in the cases of low signal to noise ratios if the inverse filter solution is used. So, the second term in the denominator of Eq. (8) solves the ill-posed problem. This means that our problem is how to evaluate the regularization parameterλ..

4. METHODS OF EVALUATING THE REGULARIZATION PARAMETER

The problem of evaluating the regularization parameter has been addressed using a diversity of techniques [6, 7]. Most of these techniques require knowledge of the noise variance of the degraded image σ^2. In many practical applications, the noise variance in the degraded image is not known and its estimation may be a tedious problem. We will survey some of these methods in the following subsections.

4.1. CONSTRAINED LEAST SQUARES (CLS) METHOD

In this method, the parameter λ is selected such that the following equation is satisfied [6]:

$$\left\|\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}(\lambda)\right\|^2 = \left\|(\mathbf{I} - \mathbf{HA}(\lambda))\mathbf{g}\right\|^2 = \|\mathbf{n}\|^2 = \varepsilon^2 = M^2\sigma^2 \quad (10)$$

Using the constrained least squares method is equivalent to assuming that the ith component of the error $\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}(\lambda)$ is Gaussian and $\left\|\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}(\lambda)\right\|^2$ is Chi-square distributed with variance σ^2 and M^2 degrees of freedom.

4.2 EQUIVALENT DEGREES OF FREEDOM (EDF) METHOD

The notation of the equivalent degrees of freedom (EDF) can also be incorporated into the evaluation of the regularization parameter [6]. The EDF method takes into account the linear dependency between the degraded image $\mathbf{g}$ and the estimated image $\hat{\mathbf{f}}(\lambda)$. Therefore, $\left\|\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}(\lambda)\right\|^2$ is Chi-square distributed with variance σ^2 and $trace(\mathbf{I} - \mathbf{HA}(\lambda))$ degrees of freedom. Thus, in this case, the constraint equation used for computing λ will be given by [6]:

$$\left\|\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}(\lambda)\right\|^2 = \left\|(\mathbf{I} - \mathbf{HA}(\lambda))\mathbf{g}\right\|^2 = \sigma^2 trace\left[\mathbf{I} - \mathbf{HA}(\lambda)\right] \quad (11)$$

4.3 MEAN SQUARE ERROR (MSE) METHOD

Another method for evaluating λ is obtained by directly minimizing the MSE function. The MSE function can be written as [6]:

$$E\left[\left\|\mathbf{e}(\lambda)\right\|^2\right] = E\left[\left\|\mathbf{f} - \hat{\mathbf{f}}(\lambda)\right\|^2\right] = \|\mathbf{f}\|^2 + E\left[\left\|\hat{\mathbf{f}}(\lambda)\right\|^2\right] - 2E\left[\mathbf{f}^t\hat{\mathbf{f}}(\lambda)\right] \quad (12)$$

Since the term $\|\mathbf{f}\|^2$ does not depend on λ, the minimization of $E\left[\left\|\mathbf{e}(\lambda)\right\|^2\right]$ is equivalent to the minimization of $E\left[\left\|\hat{\mathbf{f}}(\lambda)\right\|^2\right] - 2E\left[\mathbf{f}^t\hat{\mathbf{f}}(\lambda)\right]$. Using the Toeplitz to circulant approximation, we get [6]:

$$E\left[\left\|\hat{\mathbf{f}}(\lambda)\right\|^2\right] = E\left[\sum_{i=1}^{M^2} \frac{|h_i|^2|G_i|^2}{(|h_i|^2 + \lambda|c_i|^2)^2}\right] \quad (13)$$

and

$$E\left[\mathbf{f}^t\hat{\mathbf{f}}(\lambda)\right] = E\left[\sum_{i=1}^{M^2} \frac{|h_i|^2|F_i|^2}{(|h_i|^2 + \lambda|c_i|^2)^2}\right] = E\left[\sum_{i=1}^{M^2} \frac{|G_i|^2 - \sigma^2}{(|h_i|^2 + \lambda|c_i|^2)^2}\right] \quad (14)$$

where h_i and c_i are the Eigenvalues of $\mathbf{H}$ and $\mathbf{C}$, respectively, and F_i and G_i are the ith components

of the discrete Fourier transforms of f and g, respectively.

Substituting Eqs. (13), (14) into Eq. (12) and letting the derivative of the MSE function equal to zero we get:

$$\sum_{i=1}^{M^2} \frac{\lambda^2 |c_i|^4 |G_i|^2}{(|h_i|^2 + \lambda |c_i|^2)^3} - \sigma^2 \sum_{i=1}^{M^2} \frac{\lambda |c_i|^2}{(|h_i|^2 + \lambda |c_i|^2)^2} = 0 \quad (15)$$

In matrix form, this is equivalent to:

$$\left\| \mathbf{C}^{-1} [(\mathbf{I} - \mathbf{HA}(\lambda))]^{3/2} \mathbf{g} \right\|^2 = \sigma^2 trace[\mathbf{C}^{-2} (\mathbf{I} - \mathbf{HA}(\lambda))^2] \quad (16)$$

The solution of Eq. (16) leads to an estimate of the regularization parameter λ.

4. 4 PREDICTED MEAN SQUARE ERROR (PMSE) METHOD

Another criterion for choosing the regularization parameter is based on the minimization of the weighted error norm [6]:

$$E\left[\left\| \mathbf{He}(\lambda) \right\|^2 \right] = E\left[\left\| \mathbf{Hf} - \mathbf{H\hat{f}}(\lambda) \right\|^2 \right] \quad (17)$$

This method is based on the fact that the data points, which correspond to large Eigenvalues of H are more reliable, and thus they are weighted heavier. In matrix form, the regularization parameter λ can be evaluated by solving the following equation [6]:

$$E\left[\left\| \mathbf{He}(\lambda) \right\|^2 \right] = \left\| (\mathbf{I} - \mathbf{HA}(\lambda)) \mathbf{g} \right\|^2 + 2\sigma^2 [trace[\mathbf{HA}(\lambda)] - M^2] \quad (18)$$

4. 5 SET THEORETIC (ST) METHOD

In this method, a priori knowledge about $\mathbf{f}$ is assumed which restricts the solution to lie in a set $\mathbf{Q}_\mathbf{f}$, that is, $\mathbf{f} \in \mathbf{Q}_\mathbf{f}$ where $\mathbf{Q}_\mathbf{f}$ is an M^2 dimensional space [6].

$$\mathbf{Q}_\mathbf{f} = \left\{ \mathbf{f} \,\middle|\, \left(\left\| \mathbf{Cf} \right\|^2 \le E^2 \right) \right\} \quad (19)$$

Similarly the noise $\mathbf{n}$ is assumed to belong to a set $\mathbf{Q}_\mathbf{n}$. Since $\mathbf{n}$ must lie in a set, it follows that a given observation $\mathbf{g}$ combines with the set $\mathbf{Q}_\mathbf{n}$ to define a new set which must contain $\mathbf{f}$, i.e

$$\mathbf{Q}_{\mathbf{f/g}} = \left\{ \mathbf{f} \,\middle|\, \left(\left\| \mathbf{g} - \mathbf{Hf} \right\|^2 \le \varepsilon^2 \right) \right\} \quad (20)$$

Both of the sets $\mathbf{Q}_\mathbf{f}$ and $\mathbf{Q}_{\mathbf{f/g}}$ contain $\mathbf{f}$ and therefore $\mathbf{f}$ must lie in their intersection [6].

$$\mathbf{Q}_{\hat{\mathbf{f}}} = \mathbf{Q}_\mathbf{f} \cap \mathbf{Q}_{\mathbf{f/g}}. \quad (21)$$

To make the problem more tractable, ellipsoids are used for the sets $\mathbf{Q}_\mathbf{f}$ and $\mathbf{Q}_\mathbf{n}$. The equation of an ellipsoid is given by [6]:

$$\mathbf{Q}_\mathbf{f} = [\mathbf{f} : (\mathbf{f} - \mathbf{c}_\mathbf{f})^t \mathbf{G}_\mathbf{f}^{-1} (\mathbf{f} - \mathbf{c}_\mathbf{f}) \le 1] \quad (22)$$

where $\mathbf{c}_\mathbf{f}$ is the center of the ellipsoid and $\mathbf{G}_\mathbf{f}$ is a positive matrix, whose Eigenvalues and Eigen vectors determine, respectively, the orientation and the length of the axis of $\mathbf{Q}_\mathbf{f}$. The centers of the ellipsoids bounding the intersection of these two ellipsoids $\mathbf{Q}_\mathbf{f}$ and $\mathbf{Q}_{\mathbf{f/g}}$ are given by [23, 24]:

$$\mathbf{\hat{f}} = \left[p_1 \frac{\mathbf{H}^t \mathbf{H}}{\varepsilon^2} + p_2 \frac{\mathbf{C}^t \mathbf{C}}{E^2} \right]^{-1} . p_1 \frac{\mathbf{H}^t}{\varepsilon^2} \mathbf{g} \quad (23)$$

where $p_1 + p_2 = 1$ and $p_1, p_2 \ge 0$. Figure (1) gives a geometric interpretation of the set theoretic approach. For $p_1 = p_2$, the regularization parameter is given by [6]:

$$\lambda = (\varepsilon / E)^2 \quad (24)$$

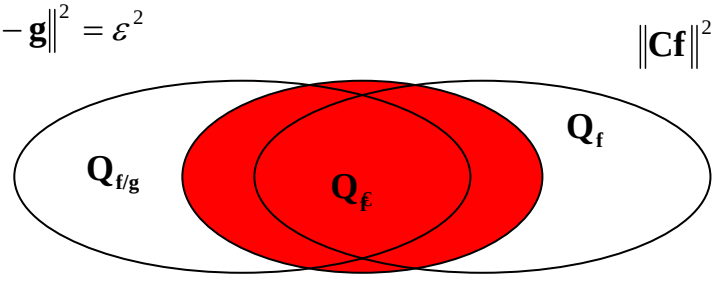


Fig. 1 – Geometric interpretation for the set theoretic approach using ellipsoids

5. THE SUGGESTED METHOD FOR EVALUATING THE REGULARIZATION PARAMETER

The different methods mentioned above for evaluating the regularization parameter yield different estimates of λ. The regularization parameter λ dictates, in general, the level of smoothness in the restored image. Thus, if two different regularization parameters λ_a and λ_b are used in restoring a degraded image with $\lambda_a > \lambda_b$, the resulting restored image $\hat{f}(\lambda_a)$ will be smoother than $\hat{f}(\lambda_b)$ [6]. The smoothness criterion doesn't reveal any thing about the MSE of the solution.

Based on the smoothness criterion, the smoothness of the solutions obtained using the different methods of evaluation differs according to the value of λ obtained with each method. Some methods yield over smoothed estimates of the restored images while others yield under smoothed estimates [6]. Accordingly, a unified approach is needed to estimate the regularization parameter λ. We will now present the proposed method for evaluating the regularization parameter that can achieve minimum mean square restoration error.

The regularized image restoration filter transfer function is expressed in Eq. (8). It is clear from this equation that the parameter λ controls the passband of this restoration filter. This is illustrated in Figs (2) to (6). Figure (2) illustrates the frequency response of a lowpass blurring operator $H(u,v)$. Figure (3) illustrates the frequency response of the inverse filter. It is clear that the inverse filter has singularities at certain frequencies. In Figs. (4) to (6), the effect of regularization is illustrated with a regularization parameter λ=0.1, 0.01 and 0.001, respectively. It is clear that the regularized filter has a bandpass nature and the parameter λ controls the location of the passband in the frequency domain.

In general, most images to be restored using this regularized restoration filter have their spatial frequencies concentrated near the low frequency region of the spectrum. When the filter passband coincides with the spectral region at which the frequency contents of the image to be restored lie, good restoration results are obtained due to the preservation of all frequency contents of the restored image. The noise frequency components, which extend to infinite frequencies lie outside the filter passband and thus are reduced.

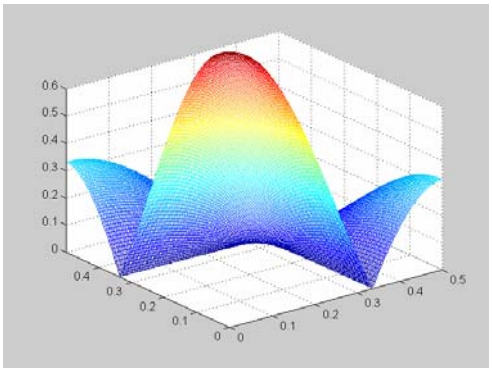


Fig. 2 – Low pass filter frequency response

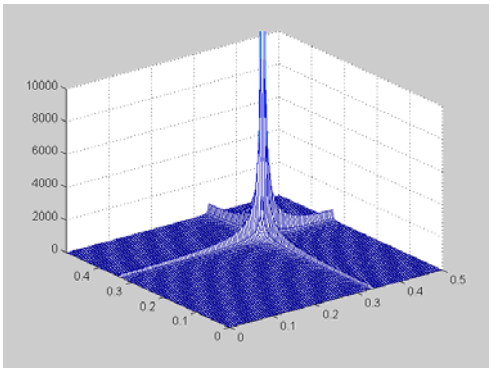


Fig. 3 – Inverse Filter frequency response

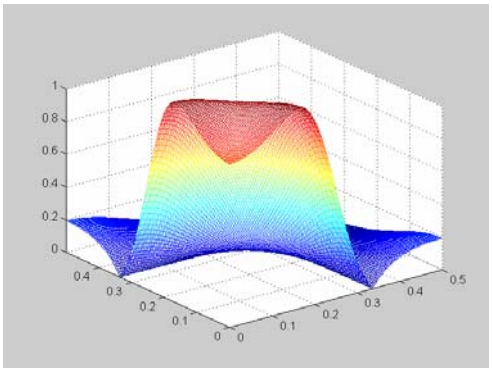


Fig. 4 – Regularized Filter frequency response λ=0.1

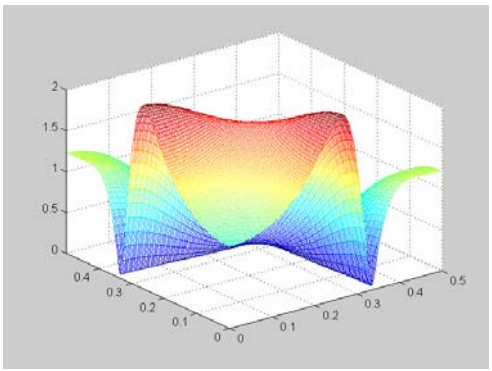


Fig. 5 – Regularized Filter frequency response λ=0.01

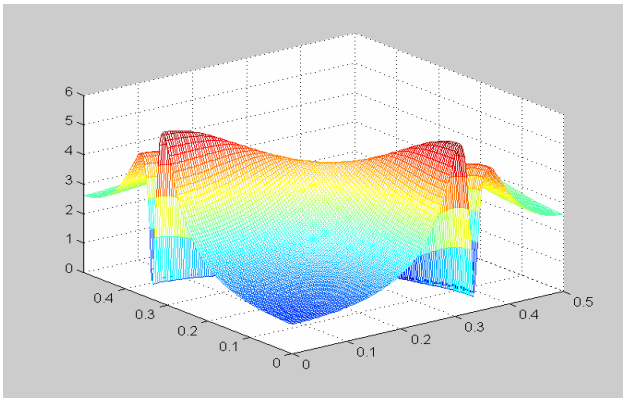


Fig. 6 – Regularized Filter frequency response λ=0.001

The suggested method for the evaluation of the regularization parameter is based on the above-mentioned concept as illustrated in Fig. (7). In this method, an initial value for the regularization parameter is assumed and the restoration process is performed on the degraded image $g(n_1, n_2)$ using the regularized filter $D(u, v, \lambda)$. The power spectrum of the restored image is estimated and then the total power in the restored image is calculated. This process is repeated iteratively by updating the regularization parameter value until a maximum value of the total power is obtained in the restored image. The value of λ which yields this maximum total power is the best estimate of λ expected to yield minimum mean square error in the restored image.

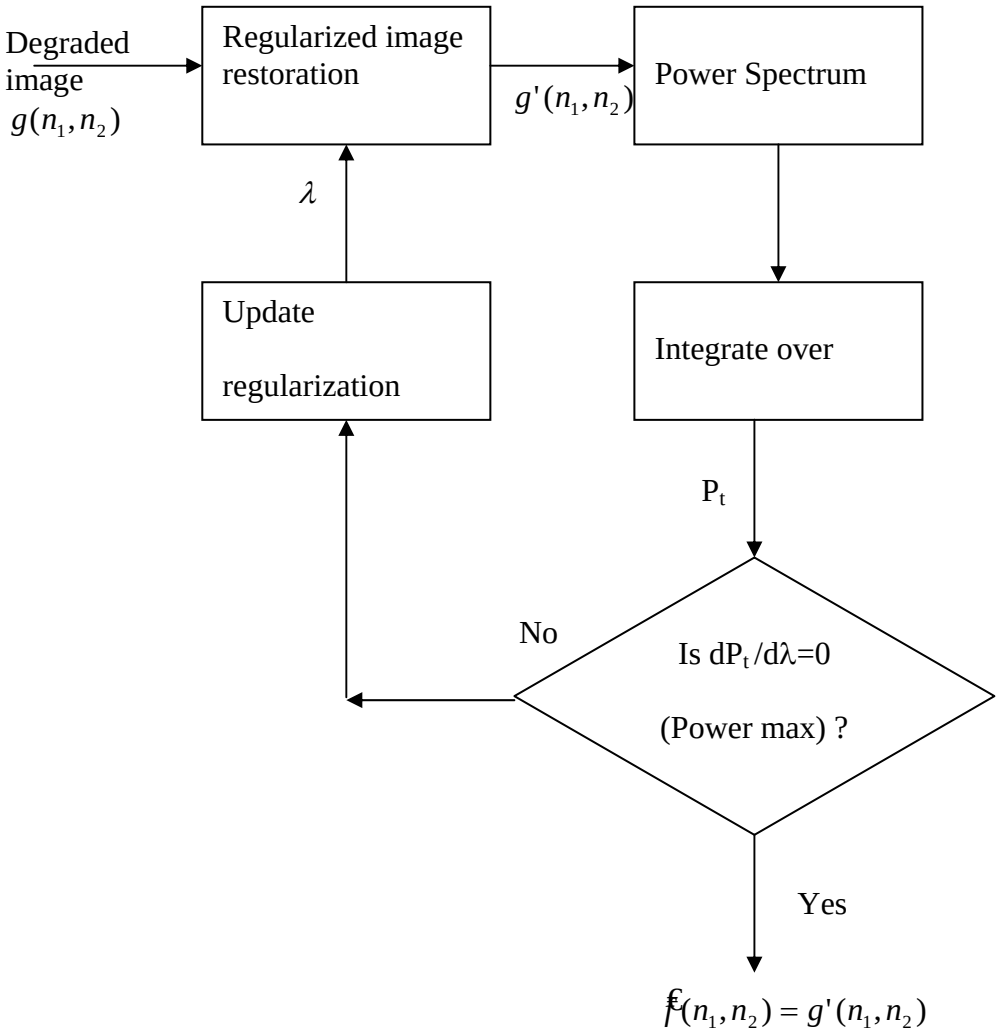


Fig. 7 – Block diagram of the suggested method for the evaluation of the regularization parameter

The auto-correlation function of the restored image $g'(n_1, n_2)$ is evaluated using the following relation:

$$R_{g'}(n_1, n_2) = \frac{1}{w^2} \sum_{k=1}^{w} \sum_{l=1}^{w} g'(k, l) g'(n_1 + k, n_2 + l) \quad (25)$$

where w is an arbitrary window length.

The power spectrum of the restored image

$g'(n_1,n_2)$ is evaluated using the following relation:

$$P_{g'}(u,v)=\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1}R_{g'}(n_1,n_2)\exp\left\{-\frac{2\pi i}{N}(n_1u+n_2v)\right\} \quad (26)$$

An integration process, which may be approximated using a summation process, is performed over the whole spectrum to evaluate the total power in the restored image $g'(n_1,n_2)$. The approximation of this integration process is performed as follows:

$$P_t=\frac{1}{N^2}\sum_{u}^{N-1}\sum_{v}^{N-1}\left|P_{g'}(u,v)\right| \quad (27)$$

When this total power is maximized, the restored image $g'(n_1,n_2)$ can be considered the best estimate of the image $f(n_1,n_2)$. If the power is not maximum, the regularization parameter is updated and the process is repeated.

6. SIMULATION RESULTS

The suggested method for evaluating the regularization parameter is tested in this section for two different images with different spatial activities; the Cameraman and the Mandrill images illustrated in Figs. (8) and (9), respectively. The steps of the proposed method illustrated in Fig. (7) are applied on both the degraded Cameraman and Mandrill images.

Fig. 8 – Original Cameraman image

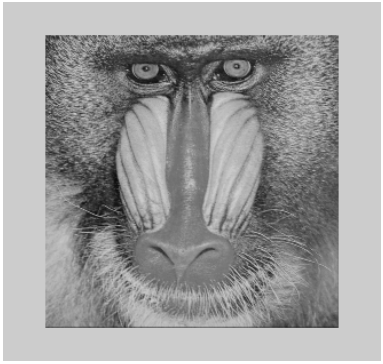


Fig. 9 – Original Mandrill image

The restoration process is performed in the presence of additive noise and a blurring operator of size 5x5 with an SNR=40 dB. The variations of the total power and MSE in the restored images with the value of the regularization parameter are plotted in Figs. (10) and (12), respectively, for the Cameraman image and in Figs. (11) and (13), respectively, for the Mandrill image. It is clear that the minimum MSE coincides with the maximum power for the Cameraman image, which is mainly an image of low frequency nature. This means that the proposed algorithm have succeeded in evaluating the best value of the regularization parameter.

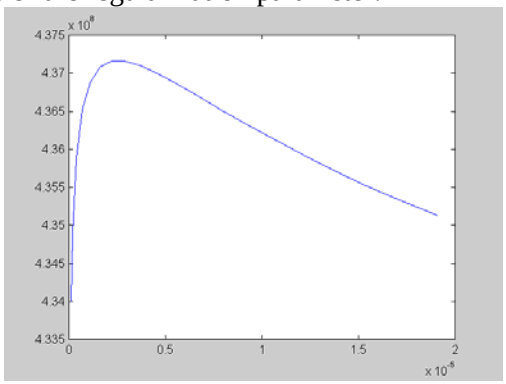


Fig. 10 – Pt versus λ for restoring the Cameraman image degraded by 5X5 operator SNR=40 dB

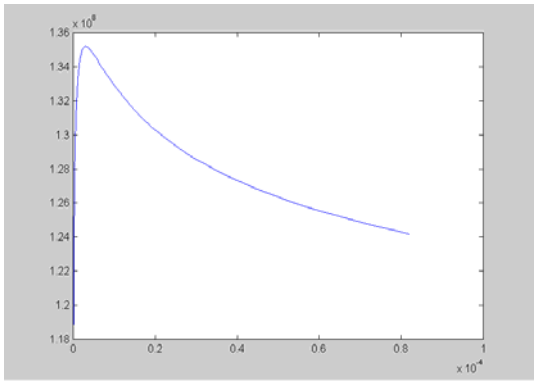


Fig. 11 – Pt versus λ for restoring the Mandrill image degraded by 5X5 operator SNR=40 dB

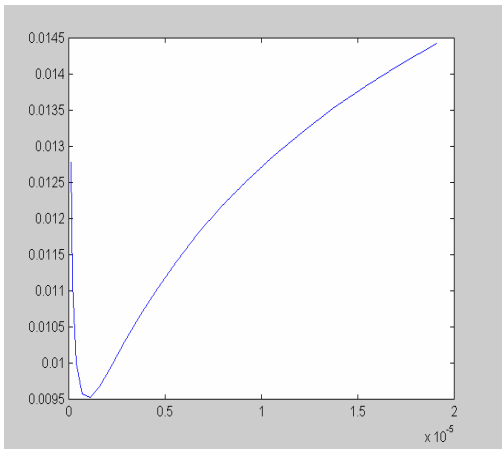


Fig. 12 – MSE versus λ for restoring the Cameraman image degraded by 5X5 operator SNR=40 dB

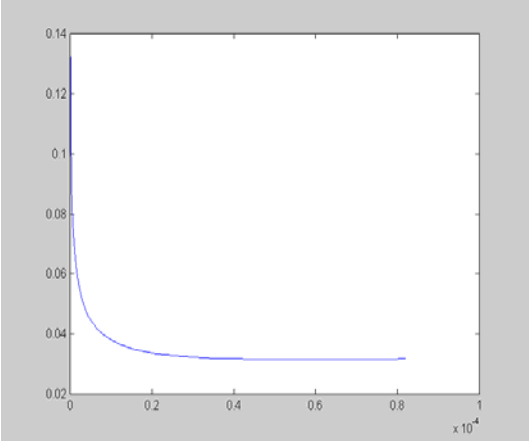


Fig. 13 – MSE versus λ for restoring the Mandrill image degraded by 5X5 operator SNR=40 dB

For the Mandrill image which is of a high frequency nature, the MSE reaches its minimum value with the maximum value of the total power in the restored image and keeps this minimum value for a long range of λ. This means that a wide range of λ can be used for the restoration of images containing much high frequency details.

7. CONCLUSIONS

This paper has presented a new iterative method for the evaluation of the regularization parameter in regularized image restoration. This method has succeeded in evaluating the value of the regularization parameter which can maximize the power in the restored image without knowledge of the noise variance in the degraded image. The proposed method can be used in the restoration of images containing either low or high frequencies. For the restoration of images with low frequencies, a single value of the regularization parameter can be used, while for the restoration of images with high frequencies, a wide margin for the regularization parameter can be used.

8. REFERENCES

[1] H.C. Anderws and B.R. Hunt, Digital Image Restoration. Englewood Cliffs, NJ: Prentice-Hall, 1977.

[2] W. K.Pratt, Digital Image Processing. John Wiley & Sons, Inc., 1991.

[3] A. K.Jain, Fundamentals of Digital Image Processing. Prentice Hall, Inc, Englewood Cliffs, N.J., USA. 1989, reprinted in 1997.

[4] R. L. Lagendijk, J. Biemond and D. E. Boekee, "Regularized Iterative Image Restoration with Ringing Reduction", IEEE Trans. Acoustics, Speech and Signal Processing, vol. 36, No.12, pp 1874-1888, Dec 1988.

[5] N. B. Karayiannis and A. N. Venetsanopoulos, "Regularization Theory in Image Restoration-the Stabilizing Functional Approach", IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 38, No.7, pp.1155-1179, July 1990.

[6] M.G. Kang and A. K. Katsagelos, "Simultaneous Iterative Image Restoration and Evaluation of the Regularization Parameter", IEEE Trans. Signal Processing, vol. 40, No.9,pp. 2329-2334 Sep. 1992.

[7] S. J. Receves and R. M. Mersereau, "Automatic Assessment of Constrained Sets in Image Restoration", IEEE Trans. Image Processing, Vol. 1, No.1,pp.119-123, Jan. 1992.

[8] N. P. Galatsanos and A. K. Katasaggelos, " Methods of Choosing the Regularization Parameter and Estimating the Noise Variance in Image Restoration and Their Relation", IEEE Trans. Image Processing, Vol. 1, No.3, pp.322-336, July 1992.

[9] A. Seghouane, "A Note on Image Restoration Using Cp and MSE" IEEE Signal Processing Letters, Vol. 15, 2008.

[10] J. Xu and S. Osher, "Iterative Regularization and Nonlinear Inverse Scale Space Applied to Wavelet-Based Denoising" IEEE Transactions on Image Processing, Vol. 16, No. 2, 2007.

[11] J. Gutiérrez, F. J. Ferri, and J. Malo, "Regularization Operators for Natural Images Based on Nonlinear Perception Models" IEEE Transactions on Image Processing, Vol. 15, No. 1, 2006.

[12] M. R. Charest, Jr. and P. Milanfar, "On Iterative Regularization and its Application", IEEE Transactions on Circuits And Systems For Video Technology, Vol. 18, NO. 3, 2008.

[13] A. E. Savakis and H. J. Trussell, "Blur Identification by Residual Spectral Matching", IEEE Trans. Image Processing, Vol. 2, No.2, pp.141-150, April 1993.

Said E. El-Khamy *has received the B.Sc. (Hons) and M.Sc. degrees from Alexandria University, Alexandria, Egypt, in 1965 and 1967 respectively, and the Ph.D. degree from the University of Massachusetts, Amherst, USA. in 1971. He joined the teaching staff of the Department of Electrical Engineering, Faculty of Engineering, Alexandria University, Alexandria, Egypt, since 1972 and was appointed as a Full-time Professor in 1982 and as the Chairman of the Electrical Engineering Department since September 2000. His Current research areas of interest include Spread-Spectrum Techniques, Mobile and Personal Communications,*

Wave Propagation in different media, Smart Antenna Arrays, Space-Time Coding, Modern Signal Processing Techniques including Neural Networks, Wavelets, Genetic Algorithms, Fractals, HOS and Fuzzy Algorithms, and their applications in Image Processing, Communication Systems, Antenna design and Wave Propagation problems. He has published about two hundreds scientific papers in national and international conferences and journals. He took part in the organization of many local and international conferences including the yearly series of NRSC (URSI) series, ISCC'95, ISCC'97, ISSPIT'2000, MELECON'2002. He also chaired technical sessions in many local and international conferences including, ISSSTA'96, Mainz, Germany, Sept. 1966; IGARSS'98, Seattle, Washington, USA, July 1998 and AP-S'99, Orlando, Florida, USA, July 1999.

Prof. El-Khamy has earned many national and international research awards among which are the Alexandria University Research Award, 1979, the IEEE, R.W.P. King best paper award of the Antennas and Propagation Society of IEEE, in 1980, the Egypt's National Engineering Research award for two times in 1980 and 1989, respectively, the Egypt's State Science & Art Decoration of the first class, 1981, the A. Schuman's-Jordan's award for Engineering Research in 1982, the Egypt's state Excellence Decoration of the first class in 1995, the IEEE Egypt' Section Award for Supervision of the Best Student Graduation Project in Communications and Networks in Egypt for the two successive years: 2000 and 2001. Recently, he received the State Scientific Excellence award in Engineering Sciences for 2002. He has also received the most cited paper award from Digital Signal Processing Journal in 2008. He is a member of the Electromagnetics Academy, a member and secretary of Egypt's national committee of URSI and is currently Egypt's National URSI Correspondent for Commission C. He is also a member of Tau Beta Pi, Eta Kappa Nu and Sigma Xi. He established the Alexandria/Egypt IEEE Subsection since 1999 and acts as its chairman since then.

***Mohiy M. Hadhoud** received the BSc and MSc degrees in Electrical Engineering from Menoufia University in Egypt in 1976 and 1981 respectively. He received the PhD degree from Southampoton University in 1987. He is currently a professor in the Dept. of Information Technology, Faculty of Computers and Information, Menoufia University. He has received the most cited paper award from Digital Signal Processing Journal in 2008. His areas of interests are signal processing, Image Processing and Digital Communications.*

***Moawad I. Dessouky** received the BSc and MSc degrees in Electrical Engineering from Menoufia University in Egypt in 1976 and 1981 respectively. He received the PhD degree from McMaster University in 1986. He is currently a professor in the Dept. Electronics and Electrical Communications, Faculty of Electronic Engineering, Menoufia University. He has received the most cited paper award from Digital Signal Processing Journal in 2008. His areas of interests are signal processing, Image Processing and Digital Communications.*

***Bassiouny M. Sallam** received the BSc degree in Electrical Engineering from Menoufia University and MSc degree from Cairo University. He received the PhD degree from Drexel University in USA. He is currently working with the Dept. Electronics and Electrical Communications, Faculty of Electronic Engineering, Menoufia University. He has received the most cited paper award from Digital Signal Processing Journal in 2008. His areas of interests are signal processing, Image Processing and Digital Communications.*

***Fathi E. Abd El-Samie** received the BSc and MSc and PhD degrees in Electrical Engineering from Menoufia University in Egypt in 1998, 2001 and 2005, respectively. He is currently a lecturer in the Dept. of Electronics and Electrical Communications, Faculty of Electronic Engineering, Menoufia University. He has received the most cited paper award from Digital Signal Processing Journal in 2008. His areas of interests are signal processing, Image enhancement, restoration, super resolution and interpolation.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

МОДЕЛЮВАННЯ ТА ОПТИМІЗАЦІЯ ПАРАЛЕЛЬНОГО ДОСТУПУ ДО ІНФОРМАЦІЇ ФАЙЛІВ БАЗ ДАНИХ ДЛЯ БАГАТОПРОЦЕСОРНИХ ЕОМ

Володимир Лісовець, Григорій Цегелик

Львівський національний університет імені Івана Франка,
вул. Університетська, 1, Львів, 79000, e-mail: kafmmsep@franko.lviv.ua

Резюме: *в статті розглянуто метод m-паралельного послідовного перегляду та два варіанти методу m-паралельного блочного пошуку, орієнтовані на їх використання в багатопроцесорних системах для пошуку інформації у файлах баз даних. Досліджено ефективність цих методів для відомих законів розподілу ймовірностей звертання до записів. За критерій ефективності взято математичне сподівання кількості паралельних порівнянь, необхідних для пошуку запису у файлі. Проведено порівняльний аналіз ефективності методів і для кожного розглянутого закону розподілу ймовірностей звертання до записів визначено свій найкращий метод. Побудовано оптимальні стратегії пошуку записів в послідовних файлах, які зберігаються у зовнішній пам'яті багатопроцесорної ЕОМ. За критерій оптимальності прийнято математичне сподівання загального часу, необхідного для пошуку запису у файлі.*

Ключові слова: *багатопроцесорні системи, математичне моделювання, паралельний пошук, бази даних.*

ВСТУП

Створення паралельних обчислювальних систем є стратегічним напрямком розвитку обчислювальної техніки. Це викликано обмеженістю максимально можливої швидкодії звичайних послідовних ЕОМ, а також наявністю обчислювальних задач, для розв'язування яких можливості існуючих засобів обчислювальної техніки недостатні.

Проблема створення високопродуктивних обчислювальних систем належить до переліку найскладніших науково-технічних задач. Організація паралельних обчислень здійснюється, в основному, за рахунок уведення надлишкових функціональних пристроїв (декількох процесорів). Якщо здійснити поділ алгоритмів, які застосовуються, на інформаційно-незалежні частини й організувати виконання кожної частини обчислень на різних процесорах, то можна прискорити процес обчислень. Такий підхід дозволяє виконувати необхідні обчислення з меншими затратами часу. Одержання максимально-можливого прискорення обмежується тільки кількістю наявних процесорів і "незалежних" частин алгоритму.

Завдяки високій надійності та продуктивності багатопроцесорні ЕОМ широко використовуються для підтримки й організації великих баз даних (БД). При розв'язуванні різноманітних задач із використанням БД основний акцент переноситься з процедур обробки інформації на процедури організації збереження та пошуку інформації в них. Тому продуктивність обчислювальних систем, орієнтованих на роботу з великими БД, у значній мірі визначається ефективністю методів паралельного пошуку інформації в БД.

1. МЕТОДИ ПАРАЛЕЛЬНОГО ПОШУКУ ТА ЇХ ЕФЕКТИВНІСТЬ

Розглядаються наступні методи паралельного пошуку записів у файлах БД для багатопроцесорних ЕОМ [1-5]:

- метод m-паралельного послідовного перегляду;
- перший варіант методу m-паралельного блочного пошуку;
- другий варіант методу m-паралельного блочного пошуку.

Проводиться аналіз ефективності цих методів для різних законів розподілу ймовірностей звертання до записів (рівномірного, "бінарного", Зіпфа й узагальненого, частковим випадком якого є розподіл, що наближено задовольняє правило "80-20" [6-9]). За критерій ефективності приймається математичне сподівання кількості паралельних порівнянь, необхідних для пошуку запису у файлі.

Зауважимо, що у випадку однопроцесорних ЕОМ ефективність методів пошуку для різних законів розподілу ймовірностей звертання до записів, а також порівняння методів за ефективністю досліджено в [10]. Деякі часткові випадки ефективності методів розглянуто в [6, 7].

Дослідження ефективності методів паралельного пошуку проведено для рівномірного розподілу ймовірностей звертання до записів і таких законів нерівномірного розподілу ймовірностей, як:

- "бінарний" розподіл

$$p_i = \frac{1}{2^i},\ i = \overline{1, N-1},\ p_N = \frac{1}{2^{N-1}},$$

де pi – ймовірність звертання до i-го запису, N – кількість записів у файлі;

- закон Зіпфа

$$p_i = \frac{1}{iH_N},\ i = \overline{1, N},$$

де $H_N = \sum_{k=1}^{N} \frac{1}{k}$;

- узагальнений закон розподілу

$$p_i = \frac{1}{i^{(c)} H_N^{(c)}},\ i = \overline{1, N},$$

де c $(0 < c < 1)$ – будь-який параметр і $H_N^{(c)} = \sum_{k=1}^{N} \frac{1}{k^c}$.

Розглянемо метод m-паралельного послідовного перегляду.

Припустимо, що до складу багатопроцесорної ЕОМ входять m процесорів, які працюють паралельно та мають спільне поле пам'яті. Пронумеруємо процесори натуральними числами від 1 до m. Суть методу m-паралельного послідовного перегляду полягає в такому. Розіб'ємо всі записи файлу умовно на блоки по m записів у кожному. Нехай $N = n \cdot m$ – кількість записів у файлі, де n – кількість блоків. Тоді при використанні m-паралельного послідовного перегляду процес пошуку запису буде складатися з низки кроків. На першому кроці i-ий процесор переглядає значення ключа i-го запису. При цьому процес перегляду може бути успішним або неуспішним. Для визначення "успішності" всі процесори повинні обмінятися інформацією. У разі успішного перегляду процес пошуку завершується. Якщо перегляд неуспішний, то на другому кроці i-ий процесор переглядає значення ключа $(m + i)$-го запису і т. д. На $(k + 1)$-му кроці (у випадку неуспішного перегляду на k-му кроці) i-ий процесор переглядає значення ключа $(km + i)$-го запису. В наслідок виконання не більше ніж n кроків шуканий запис буде знайдено, якщо він міститься у файлі. Якщо p_i – ймовірність звертання до i-го запису файла, то математичне сподівання E кількості паралельних порівнянь, необхідних для пошуку запису у файлі, виражається формулою

$$E = \sum_{i=1}^{n} \sum_{j=1}^{m} i p_{(i-1)m+j}\ .$$

Математичне сподівання у випадку різних законів розподілу ймовірностей звертання до записів має наступний вигляд [1, 2]:

- рівномірний розподіл

$$E = \frac{1}{2}\left(\frac{N}{m} + 1\right);$$

- "бінарний" розподіл

$$E = \frac{2^m}{2^m - 1};$$

- закон Зіпфа

$$E = \frac{1}{H_N}\left(H_N + \frac{N}{m} - \frac{1}{2}\ln\frac{N}{m} - C_1\right),$$

де $C_1 = 0.5 \ln 2\pi$;

- узагальнений розподіл

$$E = \frac{1}{H_N^{(c)}}\left[H_N^{(c)} + \frac{N^{1-c}}{1-c}\left(\frac{1-c}{2-c}n - \frac{\alpha^{(c)}(n)}{n^{1-c}}\right)\right],$$

де

$$\alpha^{(c)}(n) = H_n^{(c-1)} - \frac{1}{2-c}n^{2-c}.$$

Зокрема, якщо ймовірності звертання до записів задовільняють закон Зіпфа, то залежність математичного сподівання кількості паралельних порівнянь, необхідних для пошуку запису, від різної кількості процесорів показана на рис. 1.

Рис. 1 – Математичне сподівання кількості паралельних порівнянь, необхідних для пошуку запису, у випадку розподілу ймовірностей звертання до записів за законом Зіпфа, різної кількості процесорів і $N = 10^6$

На підставі порівняння ефективності методу послідовного перегляду (m=1) і методу m-паралельного послідовного перегляду доходимо висновку, що розпаралелювання методу послідовного перегляду веде до підвищення ефективності приблизно в m разів для всіх розглянутих законів розподілу ймовірностей звертання до записів, крім "бінарного".

У випадку **першого варіанту методу *m*-паралельного блочного пошуку** вважаємо, що записи впорядкованого файла розбиті на n блоків по sm записів у кожному і пошук запису у файлі здійснюємо таким чином. Спочатку шукаємо блок, який містить шуканий запис, шляхом перегляду m останніх записів блоків. Після цього пошук продовжуємо у локалізованому блоці за допомогою методу m-паралельного послідовного перегляду. Математичне сподівання E кількості паралельних порівнянь, необхідних для пошуку запису у файлі, запишемо у вигляді суми математичного сподівання кількості паралельних порівнянь, необхідних для локалізації блоку, який містить шуканий запис, і математичне сподівання кількості паралельних порівнянь, необхідних для пошуку запису в локалізованому блоці. Тоді

$$E = \sum_{k=1}^{n}\sum_{i=1}^{s}\sum_{j=1}^{m}(k+i)\,p_{(k-1)ms+(i-1)m+j}\ .$$

Математичне сподівання кількості порівнянь для розглянутих законів розподілу ймовірностей звертання до записів буде мати такий вигляд [3]:

- рівномірний розподіл

$$E = \frac{1}{2}(n+1) + \frac{1}{2}(s+1);$$

- "бінарний" розподіл

$$E = \frac{2^{ms}-s}{2^{ms}-1} + \frac{2^{m}}{2^{m}-1};$$

- закон Зіпфа

$$E = \frac{1}{H_N}\left[2H_N + n + (s-1)\left(\frac{1}{2}\ln n + C_1\right) - \frac{1}{2}\ln(ns) - C_1\right],$$

де $C_1 = 0.5\ln 2\pi$;

- узагальнений розподіл

$$E = \frac{1}{H_N^{(c)}}\left\{2H_N^{(c)} - \frac{N^{1-c}}{1-c}\left[\frac{1-c}{2-c}n - (s-1)\frac{\alpha^{(c)}(n)}{n^{1-c}} + \frac{\alpha^{(c)}(ns)}{(ns)^{1-c}}\right]\right\},$$

де

$$\alpha^{(c)}(n) = H_n^{(c-1)} - \frac{1}{2-c}n^{2-c},$$

$$\alpha^{(c)}(ns) = H_{ns}^{(c-1)} - \frac{1}{2-c}(ns)^{2-c}.$$

Якщо ймовірності звертання до записів задовільняють рівномірний закон розподілу, то залежність математичного сподівання кількості паралельних порівнянь, необхідних для пошуку запису, від різної кількості процесорів показана на рис. 2.

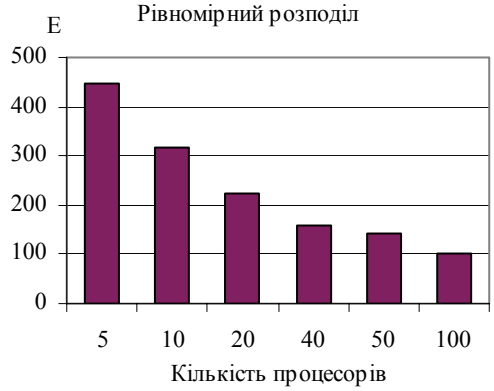


Рис. 2 – Математичне сподівання кількості паралельних порівнянь, необхідних для пошуку запису, у випадку рівномірного закону розподілу ймовірностей звертання до записів і різної кількості процесорів для $N = 10^6$

Порівнюючи ефективність методу блочного пошуку [10] та методу *m*-паралельного блочного пошуку, робимо висновок, що розпаралелювання методу блочного пошуку для всіх розглянутих законів розподілу ймовірностей звертання до записів, окрім "бінарного", суттєво підвищує ефективність. А у випадку "бінарного" закону розподілу ймовірностей збільшення кількості процесорів не підвищує ефективності роботи.

У разі **другого варіанту методу *m*-паралельного блочного пошуку** приймаємо, що записи файла розбиті на *nm* блоків по *sm* записів у кожному, тоді кількість записів у файлі буде $N = snm^2$. Пошук запису у файлі здійснюється наступним чином. Спочатку шукається блок, який містить шуканий запис, використовуючи метод *m*-паралельного послідовного перегляду серед останніх елементів блоків. Після цього пошук продовжується в локалізованому блоці також за допомогою методу *m*-паралельного послідовного перегляду. Математичне сподівання E кількості паралельних порівнянь, необхідних для пошуку запису у файлі, представимо у вигляді суми математичного сподівання кількості паралельних порівнянь, необхідних для локалізації блоку записів, і математичного сподівання кількості паралельних порівнянь, необхідних для пошуку запису в локалізованому блоці:

$$E = \sum_{k=1}^{n}\sum_{l=1}^{m} k\left(\sum_{i=1}^{s}\sum_{j=1}^{m} p_{(k-1)m^2s-(l-1)ms+(i-1)m+j}\right) + \sum_{k=1}^{nm}\sum_{i=1}^{s}\sum_{j=1}^{m} ip_{(k-1)ms+(i-1)m+j}$$

Явний вираз математичного сподівання у випадку різних законів розподілу ймовірностей звертання до записів буде мати вигляд [4]:

- рівномірний розподіл

$$E = \frac{1}{2}(n+1) + \frac{1}{2}(s+1);$$

- "бінарний" розподіл

$$E = \frac{2^{m^2s} - s}{2^{m^2s} - 1} + \frac{2^m}{2^m - 1} - \frac{s}{2^{ms} - 1};$$

- закон Зіпфа

$$E = \frac{1}{H_N}\left[2H_N + n + \frac{1}{2}s\ln(nm) - \frac{1}{2}\ln n - \frac{1}{2}\ln(nms) + C_1(s-2)\right],$$

де $C_1 = 0.5\ln 2\pi$;

- узагальнений розподіл

$$E = \frac{1}{H_N^{(c)}}\left[2H_N^{(c)} + \frac{N^{1-c}}{1-c}\left(\frac{1-c}{2-c}n + \frac{s\alpha^{(c)}(nm)}{(nm)^{1-c}} - \frac{\alpha^{(c)}(n)}{n^{1-c}} - \frac{\alpha^{(c)}(nms)}{(nms)^{1-c}}\right)\right],$$

де

$$\alpha^{(c)}(nm) = H_n^{(c-1)} - \frac{1}{2-c}(nm)^{2-c},$$

$$\alpha^{(c)}(nms) = H_{ns}^{(c-1)} - \frac{1}{2-c}(nms)^{2-c}.$$

Зокрема, якщо ймовірності звертання до записів задовільняють узагальнений закон розподілу, то залежність математичного сподівання кількості паралельних порівнянь, необхідних для пошуку запису, від різної кількості процесорів показана на рис. 3.

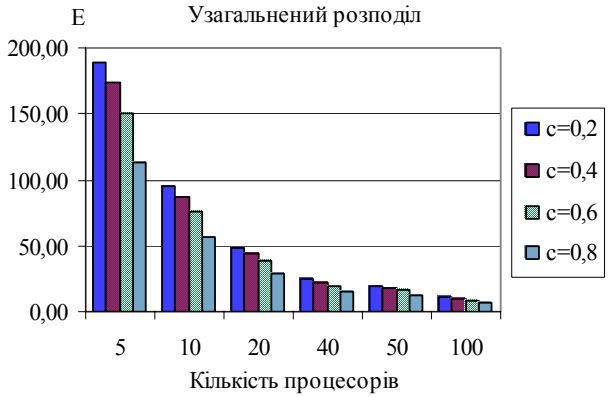


Рис. 3 – Математичне сподівання кількості паралельних порівнянь, необхідних для пошуку запису, у випадку узагальненого закону розподілу ймовірностей звертання до записів і різної кількості процесорів для $N = 10^6$

Якщо порівняти ефективність першого варіанту *m*-паралельного блочного пошуку та другого варіанту *m*-паралельного блочного пошуку, то приходимо до такого висновку: при $m = 1$ перший та другий варіанти методу дають однакову ефективність; але із зростанням кількості процесорів ефективність другого варіанту методу значно зростає порівняно з

ефективністю першого варіанту.

В результаті аналізу методів *m*-паралельного послідовного перегляду, першого та другого варіантів *m*-паралельного блочного пошуку записів у файлах БД приходимо до висновку, що обидва варіанти методу *m*-паралельного блочного пошуку є значно ефективнішими від методу *m*-паралельного послідовного перегляду, для всіх розглянутих законів розподілу ймовірності звертання до записів, крім "бінарного". Також для всіх законів розподілу ймовірностей, крім "бінарного", другий варіант методу *m*-паралельного блочного пошуку є ефективніший, ніж перший. Для розглянутих методів паралельного пошуку в табл. 1 наведені значення математичного сподівання кількості паралельних порівнянь, необхідних для пошуку запису у файлі, який містить $N = 10^6$ записів, у випадку розподілу ймовірностей звертання до записів за законом Зіпфа та різної кількості процесорів.

Таблиця 1. Математичне сподівання кількості паралельних порівнянь, необхідних для пошуку запису у файлі, для розглянутих методів паралельного пошуку у випадку розподілу ймовірностей звертання до записів за законом Зіпфа та різної кількості процесорів

	Паралельні методи пошуку		
m	Послідовний перегляд	1-й варіант блочного пошуку	2-й варіант блочного пошуку
1	69479,99	303,93	303,93
2	34740,25	211,09	152,58
4	17370,39	146,62	76,92
5	13896,42	130,39	61,80
10	6948,49	90,62	31,57
20	3474,54	63,04	16,48
40	1737,57	43,95	8,95
50	1390,18	39,16	7,46
100	695,41	27,43	4,48

2. ОПТИМАЛЬНІ СТРАТЕГІЇ

Використовуючи метод *m*-паралельного послідовного перегляду нами побудовано оптимальні стратегії паралельного пошуку інформації у послідовних упорядкованих файлах баз даних, що зберігається у зовнішній пам'яті ЕОМ, до складу якого входять *m* процесорів, які працюють паралельно і мають спільне поле пам'яті.

Припустимо, що файл, який містить *N* записів, поділений на *n* блоків, у кожному з яких є *ml* записів. Нехай $a_0 = b_0 + d_0 ml$ – час читання блоку записів в основну пам'ять, де b_0, d_0 – деякі сталі; t_0 – час виконання операції *m*-паралельного послідовного перегляду записів в основній пам'яті; p_i – ймовірність звертання до *i*-го запису файла, E_t – математичне сподівання загального часу, необхідного для пошуку запису у файлі. Приймаємо, що для пошуку запису відбувається послідовне читання блоків записів в основну пам'ять і їх *m*-паралельний послідовний перегляд. Тоді

$$E_t = \sum_{k=1}^{n}\sum_{i=1}^{l}\sum_{j=1}^{m}\{ka_0 + [(k-1)l + i]t_0\}\times \times p_{(k-1)ml+(i-1)m+j}.$$

Для E_t знайдено явний вираз у випадку розглянутих законів розподілу ймовірностей звертання до записів і визначено значення параметрів *n* і *l*, за яких математичне сподівання досягає мінімуму [5].

- Рівномірний розподіл

$$E_t = \frac{1}{2}\left[(n+1)\cdot\left(b_0 + \frac{d_0 N}{n}\right) + \left(\frac{N}{m} + 1\right)t_0\right],$$

функція E_t досягає мінімуму при

$$n = (d_0 N / b_0)^{1/2}.$$

- "Бінарний" розподіл

$$E_t = \frac{2^{ml}}{2^{ml}-1}(b_0 + d_0 ml) + \frac{2^m}{2^m - 1}t_0.$$

Для визначення параметра *l*, за якого функція E_t досягає мінімуму, отримаємо рівняння

$$2^{ml} = 1 + \left(ml + \frac{b_0}{d_0}\right)\ln 2.$$

- Закон Зіпфа

$$E_t = \frac{1}{H_N}\left[\left(H_N + n - \frac{1}{2}\ln n - C_1\right)\left(b_0 + \frac{d_0 N}{n}\right) + \left(H_N + \frac{N}{m} - \frac{1}{2}\ln\frac{N}{m} - C_1\right)t_0\right],$$

де $C_1 = 0.5\ln 2\pi$;

Для наближеного обчислення значення параметра n, за якого E_t досягає мінімуму, маємо рівняння

$$2n^2 - n = \frac{d_0 N}{b_0}\left(2H_N + 1 - \ln n - 2C_1\right).$$

- узагальнений розподіл

$$E_t = \frac{1}{H_N^{(c)}}\left\{\left[H_N^{(c)} - \frac{N^{1-c}}{1-c}\left(\frac{c-1}{2-c}n + \frac{\alpha^{(c)}(n)}{n^{1-c}}\right)\right]\left(b_0 + \frac{d_0 N}{n}\right) + \left[H_N^{(c)} + \frac{N^{1-c}}{1-c}\left(\frac{c-1}{2-c}\frac{N}{m} + \frac{\alpha^{(c)}(N/m)}{(N/m)^{1-c}}\right)\right]t_0\right\},$$

де

$$\alpha^{(c)}(n) = H_n^{(c-1)} - \frac{1}{2-c}n^{2-c}.$$

Параметр n, за якого функція E_t досягає мінімуму, можемо визначити з наступного рівняння:

$$n^{3-c} + (2-c)\cdot\left(n + \frac{2-c}{1-c}\frac{d_0}{b_0}N\right)\alpha^{(c)}(n) = (2-c)\frac{d_0}{b_0}N^c n^{1-c}H_N^{(c)} + \frac{2-c}{1-c}n\left(n + \frac{d_0}{b_0}N\right)\left(\alpha^{(c)}(n+1) - \alpha^{(c)}(n)\right).$$

Зокрема, якщо ймовірності звертання до записів задовільняють узагальнений закон розподілу, $N = 10^6$, $b_0/d_0 = 20$, то значення параметра l зображені в табл. 2.

На відміну від параметра l, оптимальні значення параметра n, за яких математичне сподівання загального часу, необхідного для пошуку запису у файлі, досягає мінімуму, не залежать від кількості процесорів, а залежать лише від зміни закону розподілу ймовірностей звертання до записів для всіх законів розподілу, крім "бінарного". Зокрема, залежність оптимального значення параметра n від зміни закону розподілу ймовірностей звертання до записів для $N = 10^6$ і $b_0/d_0 = 20$ зображено відповідно на рис. 4.

Таблиця 2. Оптимальні значення параметра *l* у випадку узагальненого закону розподілу ймовірностей та різної кількості процесорів

m	c=0.2	c=0.4	c=0.6	c=0.8
1	4201,68	3831,42	3278,69	2421,31
2	2100,84	1915,71	1639,34	1210,65
4	1050,42	957,85	819,67	605,33
5	840,34	766,28	655,74	484,26
10	420,17	383,14	327,87	242,13
20	210,08	191,57	163,93	121,07
40	105,04	95,79	81,97	60,53
50	84,03	76,63	65,57	48,43
100	42,02	38,31	32,79	24,21

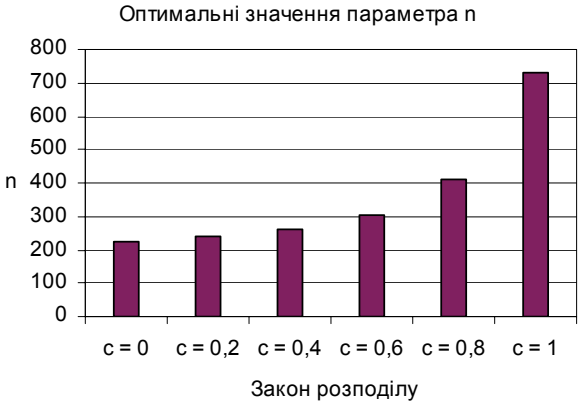


Рис. 4 – Оптимальні значення параметра n у випадку $N = 10^6$ та $b_0/d_0 = 20$ для таких законів розподілу ймовірностей звертання до записів: рівномірного (c=0), узагальненого та Зіпфа (c=1)

3. ВИСНОВКИ

Розглянуто метод m-паралельного послідовного перегляду та два варіанти методу m-паралельного блочного пошуку. Досліджено ефективність цих методів для таких законів розподілу ймовірностей звертання до записів як: рівномірний, "бінарний", Зіпфа й узагальнений, частковим випадком якого є розподіл, що наближено задовольняє правило "80-20". За критерій ефективності взято математичне сподівання кількості паралельних порівнянь, необхідних для пошуку запису у файлі. Проведено порівняльний аналіз ефективності методів і для кожного розглянутого закону розподілу ймовірностей звертання до записів визначено свій найкращий метод.

Побудовано оптимальні стратегії пошуку записів в послідовних файлах, які зберігаються у зовнішній пам'яті багатопроцесорної ЕОМ, для різних законів розподілу ймовірностей звертання до записів. За критерій ефективності взято

математичне сподівання загального часу, необхідного для пошуку запису у файлі. Визначені значення параметрів, за яких математичне сподівання досягає мінімуму.

СПИСОК ЛІТЕРАТУРИ

[1] Лісовець В. Я., Цегелик Г. Г. Метод m-паралельного послідовного перегляду записів та його використання для пошуку інформації у послідовних файлах баз даних // Фізико-математичне моделювання та інформаційні технології. – 2007. – Вип. 5. — С. 109-119.

[2] Лісовець В., Цегелик Г. Метод m-паралельного послідовного пошуку записів у файлах баз даних і його ефективність // Вісн. Львів. ун-ту. Сер. прикл. матем. та інформ. –2006. – Вип. 13.– С. 177-186.

[3] Лісовець В. Я., Цегелик Г. Г. Метод m-паралельного блочного пошуку записів у файлах баз даних та його ефективність // Відбір та обробка інформації. – 2007. – Вип. 27(103). – С. 87-92.

[4] Лісовець В. Я., Цегелик Г. Г. Один з варіантів методу m-паралельного блочного пошуку записів і його ефективність // Фізико-математичне моделювання та інформаційні технології. – 2008. – Вип. 7. – С. 103-111.

[5] Лісовець В., Цегелик Г. Моделювання та оптимізація паралельного пошуку інформації у файлах баз даних // Матеріали третьої міжнародної науково-технічна конференції: "Комп'ютерні науки та інформаційні технології" CSIT'2008 (25-27 вересня 2008р.). – Львів: Видавництво ПП "Вежа і Ко", 2008, С. 277-280

[6] Кнут Д. Искусство программирования для ЭВМ. Т. 3: Сортировка и поиск. – М.: Изд. дом "Вильямс", 2000. – 832 с.

[7] Мартин Дж. Организация баз данных в вычислительных системах. – М: Мир, 1980. – 644 с.

[8] Цегелик Г. Г. Организация и поиск информации в базах данных. – Львов: Вища шк., 1987. – 176 с.

[9] Цегелик Г.Г. Системы распределенных баз данных. – Львов: Світ, 1990, – 168с.

[10] Мельничин А. В. Моделювання та оптимізація доступу до інформації файлів баз даних. Автореферат дисертації на здобуття наукового ступеня кандидата технічних наук: спеціальність 01.05.03 "Математичне та програмне забезпечення обчислювальних машин і систем" / Мельничин А.В. – Львів: "Львівська політехніка", 2009, – 20с.

***Лісовець Володимир Ярославович**, народився 29 липня 1983 року в м. Дрогобичі Львівської області.*

В 2000 – 2005 рр навчався в Львівському національному університет (ЛНУ) ім. І. Франка на факультеті прикладної математики та інформатики. В 2005р. отримав диплом магістра.

З листопада 2007р аспірант кафедри математичного-моделювання соціально-економічних процесів факультету прикладної математики та інформатики ЛНУ ім. Франка.

***Цегелик Григорій Григорович**, народився 1942р. Закінчив Львівський державний університет, кандидат фіз.-мат наук з 1969р. Доктор фіз.-мат наук з 1990р. З 1991р. професор. З 1974р по 1984р очолював кафедру механізованої обробки економічної інформації. З 1993р по 2000р очолював кафедру обчислювальної математики, а з 2000р по теперішній час очолює кафедру математичного моделювання соціально-економічних процесів факультету прикладної математики та інформатики Львівського національного університету ім. І.Франка. Наукові інтереси: теоретичні основи інформатики та кібернетики, чисельні методи аналізу, математичне моделювання економічних процесів. Автор та співавтор понад 500 публікацій, у тому числі 3-ох індивідуальних монографій, 1-го підручника та 2-ох посібників з грифом МОН України.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

MODELING AND OPTIMIZATION OF PARALLEL INFORMATION SEARCHING IN FILES

Volodymyr Lisovets, Hryhoriy Tsehelyk

Ivan Franko National University of L'viv,
1, Universytetska str., Lviv, 79000, Ukraine,
kafmmsep@franko.lviv.ua

Abstract: *In this article the m-parallel method of sequential field searching and two variants of m-parallel block field searching method are offered. These methods are oriented to be used in multiprocessing system for information searching in files of database. We research the effectiveness of these methods for different probability distribution law of field access. The mathematical expectation of number of parallel comparisons necessary for field searching in files is taken as a criterion of effectiveness. The effectiveness of the methods is compared and analyzed. The best of offered methods is founded for every considered probability distribution. Optimal strategies of field searching in sequenced files stored in external memory of multiprocessing system are made. In this case the mathematical expectation of total time needed for field searching in files is taken as a criterion of effectiveness.*

Keywords: *multiprocessing system, mathematical modeling, parallel searching, database.*

1. THE METHODS OF PARALLEL SEARCHING AND THEIR EFFECTIVENESS

The following methods of parallel field searching in files of database for multiprocessing system are considered [1-4]:

- method of *m*-parallel sequential field searching;
- the first variant of *m*-parallel block field searching method;
- the second variant of *m*-parallel block field searching method.

We analyze the effectiveness of these methods for different probability distribution law of field access (discrete uniform, binomial, Zipf and generalized the partial occasion of witch is the probability distribution approximately satisfying the rule "80 – 20" [5-7]). The mathematical expectation of number of parallel comparisons necessary for field searching in files is taken as a criterion of effectiveness.

Let's consider the method of *m*-parallel sequential field searching.

Suppose that multiprocessing system consists of *m* processors working parallel and having common memory. Let's enumerate the processors by natural numbers from 1 to *m*. The main point of the method of *m*-parallel sequential field searching is the following. Divide conventionally all fields of file into blocks and each of them includes *m* fields. Let $N = n \cdot m$ is the number of fields in file, where *n* is the number of blocks. When method of *m*-parallel sequential field searching is used the field searching will consist of the next steps. On the first step the processor number *i* searches the value of the key of field number *i*. In this case the process of searching can be successful or failed. Each processor must exchange data with other processors on every step. In case of successful searching the field searching process is finished. In case of failed searching the processor number *i* searches the value of the key of filed number $(m + i)$ on the second step etc. If the field searching is failed on the step number *k* then on the step number $(k + 1)$ the processor number *i* searches the value of the key of filed number $(km + i)$. If the required field is located in the file then it will be found after *n* steps. If p_i is probability of access to field number *i* then mathematical expectation *E* of number of parallel comparisons necessary for field searching in files is calculated by following formula

$$E = \sum_{i=1}^{n} \sum_{j=1}^{m} i p_{(i-1)m+j} \cdot$$

In case of using the first variant of *m*-parallel block field searching method we suppose that the fields of ordered file are divided into *n* blocks each

of them includes *sm* fields. Then the filed searching will be done in this way. First of all we search the block including the needed field by looking the last *m* fields of file. After that we continue the searching in the located block by the method of *m*-parallel sequential field searching. The mathematical expectation *E* of number of parallel comparisons necessary for field searching in files lets write as the sum of mathematical expectation of number of parallel comparisons necessary for block location which includes the needed field and mathematical expectation of number of parallel comparisons necessary for field searching in the located block. Then

$$E = \sum_{k=1}^{n}\sum_{i=1}^{s}\sum_{j=1}^{m}(k+i)\,p_{(k-1)ms+(i-1)m+j}\ .$$

In case of using the second variant of *m*-parallel block field searching method let suppose that the fields of ordered file are divided into *nm* blocks and each of them includes *sm* fields. Then the number of fields of file will be $N = snm^2$. Then the field searching will be done in the following way. First of all we search the block including the needed field by using the method of *m*-parallel sequential searching among the last elements of blocks. After that we continue the searching in the located block by the method of *m*-parallel sequential searching. The mathematical expectation *E* of number of parallel comparisons necessary for field searching in files lets write as the sum of mathematical expectation of number of parallel comparisons necessary for block location which includes the needed field and mathematical expectation of number of parallel comparisons necessary for field searching in the located block:

$$E = \sum_{k=1}^{n}\sum_{l=1}^{m} k\left(\sum_{i=1}^{s}\sum_{j=1}^{m} p_{(k-1)m^2s-(l-1)ms+(i-1)m+j}\right) + \sum_{k=1}^{nm}\sum_{i=1}^{s}\sum_{j=1}^{m} ip_{(k-1)ms+(i-1)m+j}$$

2. THE OPTIMAL STRATEGIES

Using the method of *m*-parallel sequential field searching we create the optimal strategies of parallel information searching in sequenced ordered files of database that are stored in external memory of multiprocessing system. Suppose that multiprocessing system consists of *m* processors working parallel and having common memory.

Suppose that file including *N* fields is divided into *n* blocks and each of them includes *ml* fields. Let $a_0 = b_0 + d_0ml$ is the time of block field reading in the RAM, where b_0, d_0 are some constants; t_0 is the one step time of the method of m-parallel sequential field searching in the RAM; p_i is the probability of access to field number *i*, E_t is the mathematical expectation of total time needed for field searching in file. Let suppose that first of all the field blocks are read to RAM from external memory and then *m*-parallel sequential searching method is used. Then

$$E_t = \sum_{k=1}^{n}\sum_{i=1}^{l}\sum_{j=1}^{m}\{ka_0 + [(k-1)l + i]t_0\}\times \times p_{(k-1)ml+(i-1)m+j}\ .$$

3. CONCLUSION

The *m*-parallel method of sequential field searching and two variants of *m*-parallel block field searching method are considered. We researched the effectiveness of these methods for different probability distribution law of field access (discrete uniform, binomial, Zipf and generalized the partial occasion of witch is the probability distribution approximately satisfying the rule “80 – 20”). The mathematical expectation of number of parallel comparisons necessary for field searching in files was taken as a criterion of effectiveness. The effectiveness of the methods was compared and analyzed. The best of offered methods was founded for every considered probability distribution.

Optimal strategies of field searching in sequenced files stored in external memory of multiprocessing system were made for different probability distribution law of field access. In this case the mathematical expectation of total time needed for field searching in files was taken as a criterion of effectiveness. The values of parameters when mathematical expectation reaches the minimum are founded.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

TRIANGULAR DISCRETISATION FOR ANALYSIS OF MICROSTRIP MITRED BEND, BY AN ITERATIVE METHOD USING THE FAST MODAL TRANSFORM

Y. Lamhene [1)], M. Tellache [1)], B. Haraoubia [1)], H. Baudrand [2)]

[1)] Laboratoire d'Instrumentation, Faculté d'Electronique et d'Informatique
Université U.S.T.H.B. B.P 32 El Alia Bab-Ezzouar 16111 Alger,
e-mail: youcef_lamhene@yahoo.fr, haroubiab@yahoo.fr, tellachemoh@yahoo.fr
[2)] Laboratoire d'Electronique, INP, ENSEEIHT, 2, Rue Charles Camichel
31071 Toulouse Cedex, e-mail: henry.baudrand@yahoo.fr

Abstract: *In this paper, a study based on iterative method is presented. This method consists in generating a recursive relationship between a wave source and reflected waves from the discontinuity plane which is divided into cells. A high computational speed has been achieved by using Fast Modal Transform (FMT). This work is followed by an application of triangular discretization which offers several advantages over rectangular discretization. The right bend can be simulated by both rectangular and triangular cells, while the mitred bend can be exactly conformed only by the triangular mesh. Deficiencies in the rectangular approximation are identified. The computed results have been successfully compared with published data.*

Keywords: *Planar microwave circuit, mitred bend, triangular discretization, Fast Modal Transform.*

1. INTRODUCTION

Methods based on an integral formulation [1] seem to be accurate and rigorous tools for the treatment of different planar structures (Tee, Gap, Bend, Step…)

Among these methods we can distinguish methods lying in an iterative process which resolve an eigenvalue problem, and other which, by the introduction of an excitation source, reduce the equations into an inhomogeneous system via the application of the method of moments. Moreover, the integral methods [2]-[3] become efficient if basis functions have been correctly chosen.

We have developed an iterative method based on the wave concept [4-]-[5]-[6], where choice of bases functions does not arise any more.

This principle gives originality in this method. It requires, only with the precondition, a simple convergence test of the computed impedance viewed by the source.

As shown in figure 1, the planar structure, placed in a metallic box (spectral domain) is divided into cells (spatial domain) and includes three subdomains: Source, Metal and Dielectric. This concept consists in successive reflections between the circuit plane and its two sides (upper and low metallic box).

It also has alternative behaviour between space and spectral domain. This technique is combined with the Fast Modal Transform (FMT) deduced from the classic FFT (Fast Fourier Transform). Consequently a high computational speed can be achieved.

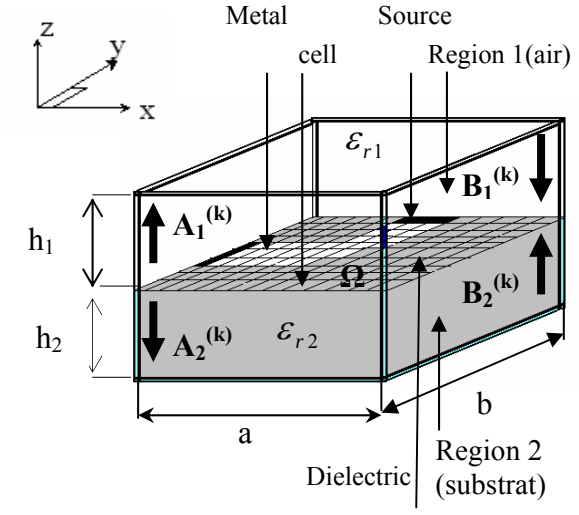


Fig. 1 – Approximate mitred bend (rectangular cells)

This method is applied to simulate microstrip corner and mitred bend using rectangular and triangular discretization.

The triangles conform exactly to any angled shapes in the discontinuity, and curved forms are reproduced in a line-segment rather than stair-case approximation using rectangular cells. In this simulation, by computing the S parameters, we show deficiencies in rectangular approximation for the mitred bend, so efficient solution requires efficient computation.

2. THEORETICAL DEVELOPMENTS

2.1 PRINCIPLE

The implementation of the iterative process consists of establishing a recursive for the relationship between the waves in the two regions 1 and 2, using the reflection in spectral domain (equation 1) and the diffraction (with boundary conditions required on Ω) in the spatial domain (equation 2).

The governing equations employed are:

$$B_i^{(k)} = \Gamma A_i^{(k-1)} \tag{1}$$

$$A_i^{(k)} = S_{\text{int}} B_i^{(k)} + A_i^{(0)} \tag{2}$$

k : is the iteration number

Bi **:** is the reflected wave on region 1or 2.

Γ : is a reflection operator in the spectral domain.

Ai : is the incidental wave on region 1 or 2.

S_{int} : is a diffraction operator on the discontinuity plane Ω in the space domain.

$A_i^{(0)} = \dfrac{\vec{E}_0}{\sqrt{Z_{0i}}}$: is the source wave (initial wave).

$\vec{E}_0$: is the electric field produced by the source.

Z_{0i} : is the intrinsic impedance of region 1 or 2.

In the equation (2), *Bi* becomes incidental wave and *Ai* the reflected one.

The iterative process is given in figure 2, for one iteration.

It uses the FMT (Fast Modal Transform) deduced from the FFT (Fast Fourier Transform) which makes it possible to accelerate the digital processing on the whole of the pixels of the planar considered circuit.

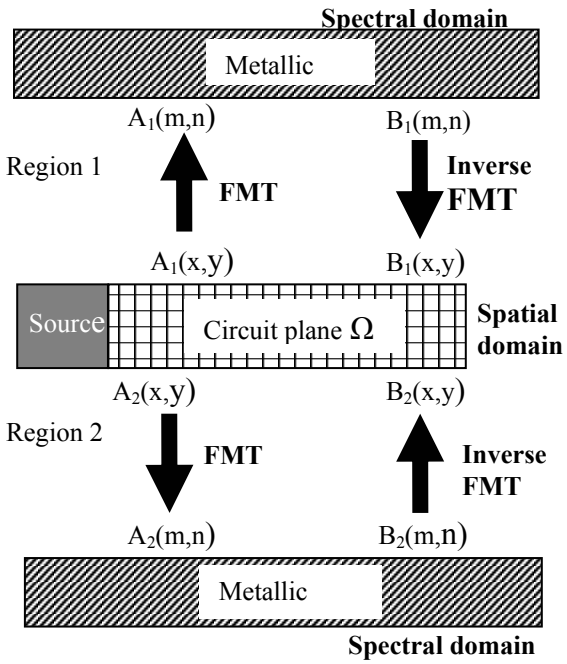


Fig. 2 – Fast Modal Transform for one iteration

2.2. REFLECTION IN THE SPECTRAL DOMAIN

The reflection operator is expressed as:

$$\Gamma = \frac{1 - Z_{0i} Y}{1 + Z_{0i} Y} \tag{3}$$

Y is the operator admittance

It uses the bases functions $\left| f_{mn}^{\alpha} \right\rangle$ of the box modes as follows:

$$Y = \sum_{m,n} \left| f_{mn}^{\alpha} \right\rangle Y_{mn}^{\alpha} \left\langle f_{mn}^{\alpha} \right| \tag{4}$$

α is the TE or TM mode.

Γ can be expressed in another manner according to the formalism of the mathematical operators [7]:

$$\Gamma = \sum_{m,n} \left| f_{mn}^{TE} \right\rangle \Gamma_{mn}^{TE} \left\langle f_{mn}^{TE} \right| + \sum_{m,n} \left| f_{mn}^{TM} \right\rangle \Gamma_{mn}^{TM} \left\langle f_{mn}^{TM} \right| \tag{5}$$

2.3. REFLECTION AND DIFFRACTION IN THE SPACE DOMAIN

The wave concept is introduced by writing the electric field E_i and the current density J_i in terms of waves.

It leads to the following set of equations [5]

$$\vec{E}_i = \sqrt{Z_{0i}}\left[\vec{A}_i + \vec{B}_i\right] \quad (6)$$

$$\vec{J}_i = \frac{1}{\sqrt{Z_{0i}}}\left[\vec{A}_i - \vec{B}_i\right] \quad (7)$$

Where: $$Z_{oi} = \sqrt{\frac{\mu_0}{\varepsilon_0 \varepsilon_{ri}}} \quad (8)$$

The diffraction operator $\mathcal{S}_{\text{int}}$ is defined in the space domain. It gives the boundary conditions and the relations of continuity of the tangential fields on the interface Ω

The planar circuit is divided to three subdomains (metal, dielectric and source)

2.3.1 METALLIC SUBDOMAIN

The boundary condition on the metal can be written as:

$$\begin{cases} \vec{E}_1 = \vec{E}_2 = 0 \\ \vec{J}_1 = \vec{J}_2 = \vec{J}_0 \end{cases} \quad (9)$$

Considering calculation by wave concept, we have:

$$\sqrt{Z_{01}}\left(\vec{A}_1 + \vec{B}_1\right) = \sqrt{Z_{02}}\left(\vec{A}_2 + \vec{B}_2\right) = 0$$

$$\Rightarrow \begin{cases} \vec{A}_1 = -\vec{B}_1 \\ \vec{A}_2 = -\vec{B}_2 \end{cases} \quad (10)$$

These two relations can be introduced in the following matrix form as:

$$\begin{bmatrix} \vec{A}_1 \\ \vec{A}_2 \end{bmatrix} = \begin{pmatrix} -H_m & 0 \\ 0 & -H_m \end{pmatrix} \begin{pmatrix} \vec{B}_1 \\ \vec{B}_2 \end{pmatrix} \quad (11)$$

Where H_m is the indicating function of the metal region which is defined as:

$$H_m = \begin{cases} 1 & \text{on the metal} \\ 0 & \text{elsewhere} \end{cases} \quad (12)$$

$$\Rightarrow \quad \mathcal{S}_{\text{int}\,m} = \begin{bmatrix} -H_m & 0 \\ 0 & -H_m \end{bmatrix} \quad (13)$$

The waves are completely reflected by metal. Thus: $\mathcal{S}_{\text{int}\,metal} = -1$

2.3.2 DIELECTRIC SUBDOMAIN

The boundary and the continuity conditions on the dielectric are:

$$\begin{cases} \vec{J}_1 + \vec{J}_2 = 0 \\ \vec{E}_1 = \vec{E}_2 \neq 0 \end{cases} \quad (14)$$

Considering the waves, we have:

$$\begin{bmatrix} \vec{A}_1 \\ \vec{A}_2 \end{bmatrix} = \begin{bmatrix} \frac{1-N^2}{1+N^2}H_i & \frac{2N}{1+N^2}H_i \\ \frac{2N}{1+N^2}H_i & -\frac{1-N^2}{1+N^2}H_i \end{bmatrix} \begin{bmatrix} \vec{B}_1 \\ \vec{B}_2 \end{bmatrix} \quad (15)$$

Where N is: $$N = \sqrt{\frac{Z_{01}}{Z_{02}}} \quad (16)$$

Hi is the indicating function of the dielectric such as:

$$H_i = \begin{cases} 1 & \text{on the dielectric} \\ 0 & \text{elsewhere} \end{cases} \quad (17)$$

We deduce the diffraction operator $\mathcal{S}_{\text{int}}$ in this subdomain:

$$\mathcal{S}_{\text{int}\,d} = \begin{bmatrix} \frac{1-N^2}{1+N^2}H_i & \frac{2N}{1+N^2}H_i \\ \frac{2N}{1+N^2}H_i & -\frac{1-N^2}{1+N^2}H_i \end{bmatrix} \quad (18)$$

2.3.3 SOURCE DOMAIN

In the iterative method, we use a cell source for exciting a planar circuit. In the figure 3, we show a cell source obtained by a classic source of tension.

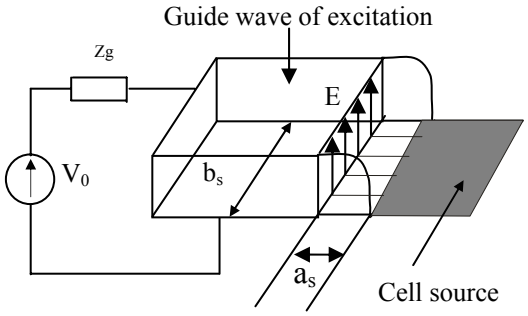


Fig. 3 – Planar source

The source emits energy through surface:

$$S_s = a_s b_s$$

This energy is guided in a waveguide and the source takes its energy of an external source of power with potential V_0 and internal impedance Z_g

The equivalent diagram of such a source (excitation from the region 1) is given in figure 4.

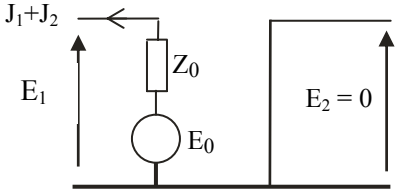


Fig. 4 – Source equivalent diagram

The boundary condition can be written as:

$$\begin{cases}\vec{E}_1 = \vec{E}_0 - Z_0\left(\vec{J}_1 + \vec{J}_2\right) \\ \vec{E}_2 = 0 = \sqrt{Z_{02}}\left(\vec{A}_2 + \vec{B}_2\right) \Rightarrow \vec{A}_2 = -\vec{B}_2\end{cases} \quad (19)$$

Thus, the reflected waves in the two regions can be obtained:

$$\begin{bmatrix}\vec{A}_1 \\ \vec{A}_2\end{bmatrix} = \frac{1}{n_1+1}\begin{bmatrix}n_1 - 1 & \frac{2Z_0}{\sqrt{Z_{01}Z_{02}}} \\ 0 & -(n_1+1)\end{bmatrix}\begin{bmatrix}\vec{B}_1 \\ \vec{B}_2\end{bmatrix} + \frac{1}{n_1+1}\begin{bmatrix}\frac{E_0}{\sqrt{Z_{01}}} \\ 0\end{bmatrix} \quad (20)$$

With:
$$n_1 = \frac{Z_0}{Z_{01}} \quad (21)$$

Z_0: is the characteristic impedance.

Finally:

$$\mathbb{S}_{\text{int}.S} = \frac{1}{n_1+1}\begin{bmatrix}n_1 - 1 & \frac{2Z_0}{\sqrt{Z_{01}Z_{02}}} \\ 0 & -(n_1+1)\end{bmatrix} \quad (22)$$

In conclusion, the computation of the waves reflected by the discontinuity plane Ω of the structure for the three subdomains is given by the equation:

$$\begin{bmatrix}A_1 \\ A_2\end{bmatrix} = \left[\mathbb{S}_{\text{int}.m} + \mathbb{S}_{\text{int}.d} + \mathbb{S}_{\text{int}.s}\right]\begin{bmatrix}B_1 \\ B_2\end{bmatrix} + \begin{bmatrix}A_{01}\, H_s \\ A_{02}\, H_s\end{bmatrix} \quad (23)$$

H_S: is the indicating function of the source domain:

$$H_S = \begin{cases}1 & \text{on the sources} \\ 0 & \text{elsewhere}\end{cases} \quad (24)$$

3. CONTRIBUTION OF FAST FOURIER TRANSFORM

The Fast Modal Transform procedure (figure 2) is based on the Fast Fourier Transform.

It takes into account the modes of the case (spectral domain). This transform gives the possibility to pass quickly from the (x,y) space domain to (m,n) modal or spectral domain and vice versa (opposite FMT) [6]:

$$\vec{\Phi}(x,y) \rightarrow \vec{\Phi}(m,n) \quad (25)$$

$$\vec{\Pi}(x,y) = \sum_m \sum_n a_{mn}^{TE}\, \vec{f}_{mn}^{TE}(x,y) + \sum_m \sum_n a_{mn}^{TM}\, \vec{f}_{mn}^{TM}(x,y) \quad (26)$$

a_{mn}^{TE} and a_{mn}^{TM} are the amplitudes of TE and TM modes.

$$a_{mn}^{TE} = \left\langle \vec{f}_{mn}^{TE}(x,y) \middle| \vec{\Pi}(x,y) \right\rangle \quad (27)$$

$$a_{mn}^{TM} = \left\langle \vec{f}_{mn}^{TM}(x,y) \middle| \vec{\Pi}(x,y) \right\rangle \quad (28)$$

4. PASSAGE FROM RECTANGULAR TO TRIANGULAR CELLS

The software of triangular cells is deduced from the program of the rectangular cells (used for the right bend and the approximated bend).

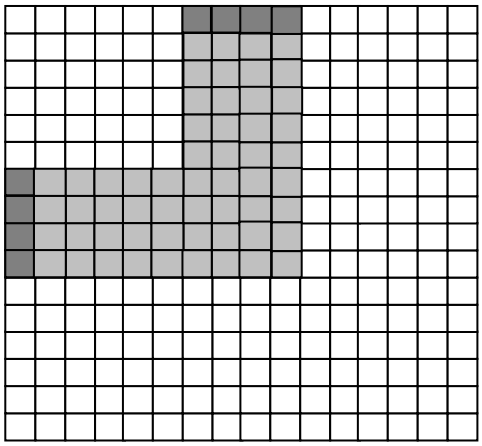


Fig. 5 – Right bend

It is necessary to use this modified software for the planar circuits having geometrical forms adapted to these cells.

Indeed, each rectangular cell gives us two triangular cells having the same surface, but different position of the centre of gravity. They are also two types of triangular cells differentiated by their slope which is positive or negative.

For the discretization of a zone of the circuit, the program requires to choose between the cell with positive or negative slope and surface in a number of cells in the two grids (positive grid or negative).

Thus, for any grid, the cells of the plane circuit are codified:

0 : for “dielectric” cells
1 : for “metal” cells
2 : for “source” cells

An example for a grid of 16x12 triangular cells (with positive slope) is given in figure 6.

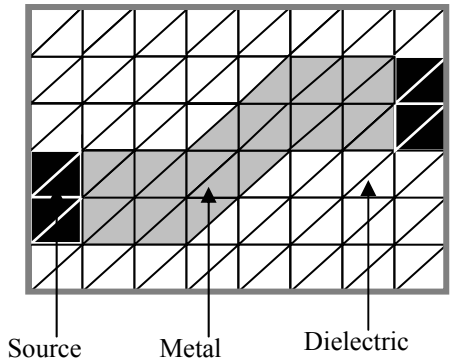


Fig. 6 – Triangular cells with positive slope

The corresponding coded grid is:

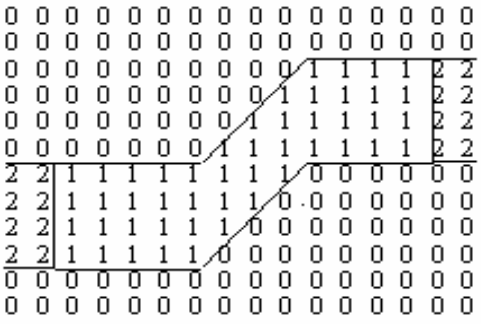


Fig. 7 – Coded grid of 16x12 triangular cells

The rectangular discretization software is applied, to the approximated mitred bends (rectangular cells).

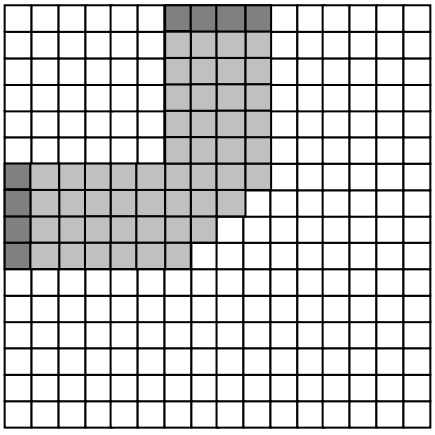


Fig. 8 – Staircase approximation mitred bend (rectangular cells)

We obtain representation with positive grid of the true mitred bend indicated by figure 9.

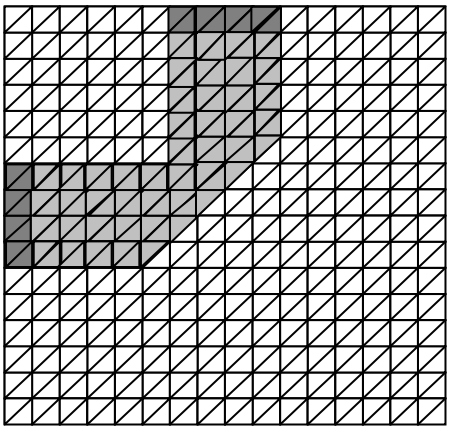


Fig. 9 – Mitred bend (triangular cells)

5. RESULTS

The dimensions of the studied bend are :

W = 0.75mm (thickness of the ribbon).

The surface of Ω plan is: S=8mm x 8mm (a=b).

The permittivity of the substrate is : εr = 10.

The total number of cells of the plane circuit is 32x32 rectangular cells, then evidently, 64x64 triangular cells.

The treatment starts with the convergence test of the input admittance viewed by the source exciting the studied bend.

This is shown in figure 10. We see that, it is enough, approximately 300 iterations, to ensure the convergence of the iterative method for modelling the planar bend. The real part of admittance which is null, correctly show the fact that the two extremities of the bend are in contact with the metallic box (short-circuit) [10].

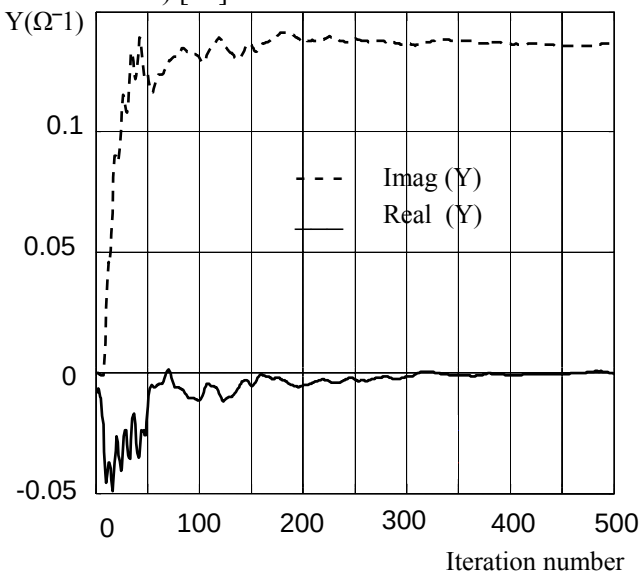


Fig. 10 – Admittance viewed by the source

With a frequency, f=14 GHz, and for 300 iterations we have determined the current density J_i of the mitred bend, which is represented in figure 11 for a plan of surface: 32x32 rectangular cells, or 64x64 triangular cells.

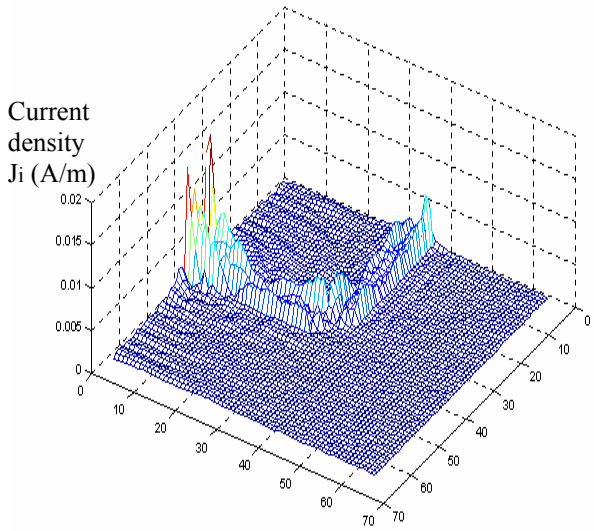


Fig. 11 – Current density J_i of the mitred bend

Finally, we have calculated the S parameters, particularly the reflection parameter S_{11} for the three cases of planar bend: right bend (without mitre), staircase approximation mitred bend (rectangular cells), true mitred bend (triangular cells).

Figure 12 shows a comparison of the results between FMT iterative method and the method of moments.

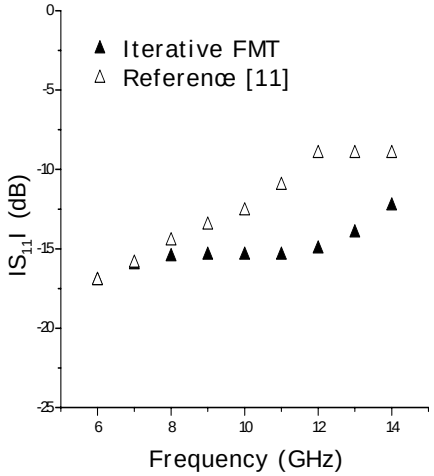


Fig. 12 – S_{11} Parameters of the right bend

We notice a considerable variation for frequencies higher than 8 GHz (dependent on dimensions of the planar bend).

Indeed, the right bend is very dispersive for the high frequencies, from where it's necessary to use mitred bend. In addition, towards 12 GHz we have a resonance frequency which requires a consequent increase in the iteration number to ensure a good convergence of the method.

This same parameter is calculated for the Approximated mitred bend and the results are given in figure 13. In this case, the effect of discontinuity of the bend is reduced by the approximated mitred bend and the results are improved.

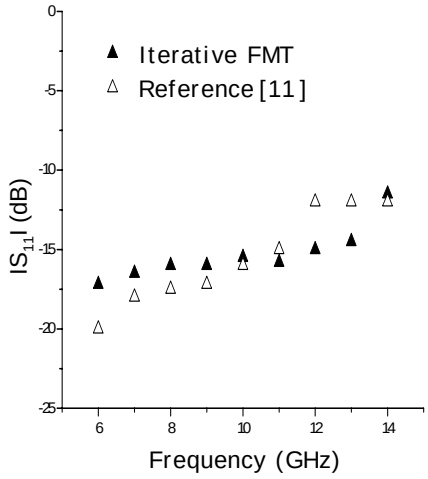


Fig. 13 – S_{11} Parameters of the Approximated mitred bend

Lastly, the software using the triangular cells is carried out, and gives this parameter for the true mitred bend, shown in figure14. The real mitred bend, obtained by the triangular cells, gives the best results, even for the resonance frequency where the iteration number is maintained to 300.

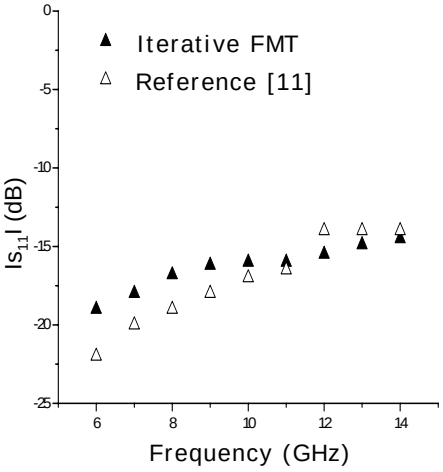


Fig. 14 – S_{11} Parameters of the mitred bend

We finish by doing a comparison, between the three cases of this planar bend. The figure 15 show the best transmission of the signal in the case of true mitred bend, where the reflection parameter is weaker, particularly in the [6 GHz ;8 GHz).

In this band, with regard to the rectangular discretization, the triangular one gives an average attenuation of 10.3 percent for the reflection on this mitred bend.

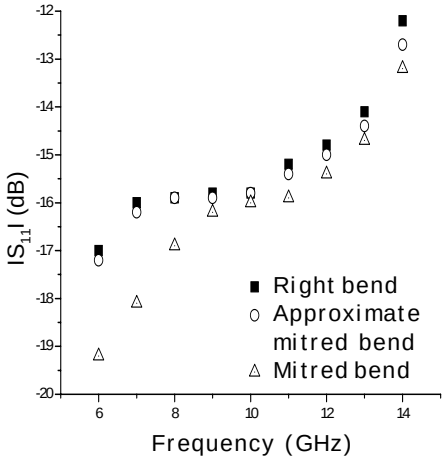


Fig. 15 – Comparison of S_{11} parameter between the three cases of bend.

6. CONCLUSION

We described in this article, a recent method integral, namely the iterative method using a Fast Modal Transform which contributes to the numerical processing power. This new concept is exempted of the delicate choice of bases functions which exist in the other integral methods.

Our results were validated by comparison with work of R. KIPP and C H. CHAN [11]. They have used the method of moments [1]. We have showed that a judicious choice of the shape of the cells, permits to treat planar structure, such a complex is it, with very good results.

7. REFERENCES

[1] R.F. Harrington**,** Field Computation by Moment Methods. New York: Macmillan, 1968.

[2] J. R. Mosig and F. E. Gardiol**,** "Integral equation techniques for the dynamic analysis of microstrip discontinuities", Alta Frequenza, vol. 57, no. 5, pp. 171-181, June 1988.

[3] P. B. Katehi and N. G. Alexopoulos, "Frequency-dependent characteristics of microstrip discontinuities in millimeter-wave integrated circuits", IEEE Trans. Microwave Theory Tech., vol. MTT-33, pp. 1029-1035, Oct. 1985.

[4] H. Baudrand,"Representation by equivalent circuit of the integral methods in microwave passive elements", Eumc, Budapest, Sept.1990.

[5] M. Azizi, H. Baudrand**,** "A new itérative method for scattering problems", European Microwave Conf., Proc., Vol. 1, pp. 255-258, Bologne, Italy, 1995.

[6] Y. Lamhene, R. Garcia, H. Baudrand, "Modélisation d'une antenne patch par une méthode itérative basée sur le concept d'ondes", OHD99, Besançon September 1999.

[7] Y. Lamhene, A. Khoda, H. Baudrand, B. Haraoubia,"Finline Structure Characterization by the Transverse Resonance Method (T.R.M)", *International Symposium on Signals Circuits and Systems (SCS'97), pp142-145, Roumania, October 1997*.

[8] A. Khodja, Y. Lamhene, H. Baudrand, B. Haraoubia, "Convergence Study of Cutoff Frequencies of the Finline Structure for Different Types of Trial Function by the Transverse Resonance Method", *International Conference on Computational Electromagnetics and Its Applications (ICCEA'99), Beijing, China, November 1999*.

[9] P. Anders and F. Arndt, "Microstrip discontinuity capacitances and inductances for double steps, mitered bends with arbitrary angle, and asymmetric right-angle bends", IEEE Trans Microwave Theory Tech.,vol. MTT-28, pp. 1213-1217, Nov. 1980.

[10] A.A.Omar, Y.L. Chow,L.Roy and M.G.Stubbs "Effects of air-bridges and mitering on coplanar waveguide 90° bends: theory and experiment", IEEE MTT-S Digest, 1993, pp. 823-826.

[11] R. Kipp and C. H. Chan, "Triangular-Domain Basis Functions for Full-Wave Analysis of Microstrip Discontinuities", IEEE Trans. Microwave Theory Tech., vol. MTT-41, no. 6/7, pp. 1187-1194, June/July 1993.

Youssef Lamhene *received bachelor's degree then master's degree in Electronic, Electrotechnic and Automatic (EEA), in 1978 and 1980 from Lille University in France, and later on, in the same place, the DEA diploma in 1982 and the third cycle Doctorate diploma in Biotechnology domain in 1984. Since 1986 he teaches graduate level courses in applied electronic,*

microwave and antenna theory at the sciences and technology university of Algiers, in Algeria. His research interests are in the area of electromagnetic simulations of microwave and millimeter-wave passive components.

Mohamed Tellache *was born in Draa El Mizane, Algeria in 1960. He received the Engineer degree in Electrical Engineering from the Ecole Nationale Polytechnique of Algiers in 1985. He received his Master degree in 1988 in electronic communication from Ottawa University, Ottawa, Canada. During 1995/1997, Mr. Tellache spent eighteen months at the laboratory of electromagnetics-Toulouse, France working principally on microwaves technology and modeling. Presently he is Assistant Professor of Electrical Engineering and Adjoin Dean at the faculty of Electronics and Computers. His present research interests focus on microwaves circuits, modeling and applied mathematics. Mr. Tellache is author and co-author of more than 15 communications in the field.*

Brahim Haraoubia *was born in Taoura, Algeria in 1957. He received the Engineer degree in Electrical Engineering from the Ecole Nationale Polytechnique of Algiers in 1981. He received his Master and PhD degrees in 1982 and 1988 both in electronic communication from Université de Rennes, France. Presently he is full Professor of Electrical Engineering at the faculty of Electronics and Computers. His present research interests focus on microwaves circuits and modeling VHF/UHF circuits. Pr. Haraoubia is author and co-author of more than 30 publications and communications in the field and has deposit five patents in his field.*

Henri Baudrand *is Emeritus Professor at ENSEEIHT, Toulouse, France. He specializes in modeling passive and active circuits and antennas. He is the author and co-author of three books at Cepadues edition. He has co-authored over 100 publications in journals and 250 contributions to international conferences. He is a follow member of IEEE society, a member of "Electromagnetism Academy" and a senior member of IEE society. He was the president of URSI, France commission B for six years (1993 to 1999) and the president of the IEEE-MTT-ED French chapter. He was awarded "officier des palmes académiques", and Director honoris causa of Lasi University.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

ГИБРИДНЫЕ ИНФОРМАЦИОННЫЕ МОДЕЛИ В СИСТЕМАХ ОБРАБОТКИ ИЗОБРАЖЕНИЙ

С. Антощук, О. Бабилунга

Кафедра информационных систем, Институт компьютерных систем,
Одесский национальный политехнический университет,
проспект Шевченко, 1, г. Одесса, 65044, Украина,
svetlana_onpu@mail.ru, babilunga@mail.ru

Резюме: *в данной статье проанализированы информационные процессы, происходящие в системах обработки визуальной информации. Представлен подход к построению гибридных информационных моделей в системах обработки изображений, который позволил создавать более эффективные методы и информационные технологии обработки, анализа и распознавания изображений.*

Ключевые слова: *информационные технологии, визуальная информация, обработка и распознавание изображений, гибридные модели, контурный анализ.*

ВВЕДЕНИЕ

В современных информационно-управляющих системах (ИУС) все чаще применяют системы обработки визуальной информации (СОВИ), предназначенные для решения самых разнообразных задач: измерение и контроль линейных и угловых размеров неподвижных и движущихся объектов, счет объектов и учет продукции, контроль формы изделий и определение отклонений от эталонных форм и др. [1, 2]. Особенно эффективно использование СОВИ в тех случаях, когда непосредственное взаимодействие с объектом контроля невозможно. Основное назначение таких систем – отображение, обработка и устранение избыточности визуальной информации, представление ее в таком виде и количестве, которое позволяет выделить существенные характеристики систем, объектов, процессов, оценить их состояние, выработать в случае необходимости сигналы управления процессом, объектом [3, 4]. Визуальная информация (ВИ) обрабатывается практически во всех диапазонах электромагнитных и акустических волн: оптическом, радио-, рентгеновском, ультразвуковом и др. В настоящее время область применения СОВИ ограничена из-за их недостаточной эффективности и универсальности, поэтому повышение качества обработки визуальной информации, представленной в виде изображений, является важной научной и практической проблемой. Ее решение возможно при эффективном моделировании СОВИ и происходящих в них процессов [3].

1. АНАЛИЗ ИНФОРМАЦИОННЫХ ПРОЦЕССОВ В СИСТЕМАХ ОБРАБОТКИ ИЗОБРАЖЕНИЙ

Поскольку СОВИ относятся к сложным системам, работающим в условиях неопределенности исходной информации, основной функцией которых является извлечение существенной информации, определяющей эффективность работы ИУС в целом, то для их моделирования целесообразно использовать системный подход. Иерархия целей применения СОВИ отражена в системной модели (рис.1). Полученная ВИ в ИУС может быть предназначена для наблюдения за объектом (сбор, преобразование и первичная обработка данных, визуализация), обнаружения объекта или измерения информации о нем (контроль, обнаружение отклонений) и опознавания объекта или его характеристик (диагностика, оценка параметров объекта управления и классификационное заключение) [5].

Рис. 1 – Системная модель применения СОВИ в ИУС

Существующие СОВИ включают системы формирования изображений и автоматизированные системы обработки изображений (СОИ). Системы формирования изображений осуществляют восприятие и фиксацию ВИ на носителе информации. Выбор конкретного типа считывания визуальной информации производится на основании известных (точно или приближенно) характеристик исследуемых объектов с учетом необходимых эксплуатационных требований.

Поскольку роль СОИ, в соответствии с приведенной системной моделью, может быть разной, то при их создании целесообразно учитывать модели обработки изображений, модели входных и выходных данных на разных уровнях обработки. Такой подход соответствует информационной теории, в рамках которой рассматриваются три тесно связанные между собой проблемы – входное представление, обработка (преобразование) и выходное представление изображений [4]. Они учитывались при разработке моделей.

Анализ основных информационных подходов к моделированию процессов представления и обработки ВИ в СОИ с учетом системной модели позволяет выделить четыре уровня глубины обработки ВИ (табл. 1). В таблице введены следующие обозначения: $B(x,y)$ – яркостное представление ВИ; $I'(x,y)$ – яркостное представление ВИ в виде изображения; $\Phi\{\cdot\}$ – оператор обработки; $I_j(x,y)$ – сегментированная область изображения; $KP_j(x,y)$ – контур области изображения; $\mathbf{K}_{ji}$ – вектор контурного описания области изображения; $\mathbf{T}_j$ – вектор характерных точек; $\mathbf{C}_j$ – вектор признаков; Q_L – классификационное решение (совокупность выделенных классов).

Таблица 1. Уровни глубины обработки визуальной информации

Уровень обработки ВИ	Характер анализа изображения	Модель информационного процесса [4]
1	Предварительная обработка	$B(x,y) \rightarrow \Phi\{I'(x,y)\} \rightarrow I(x,y)$
2	Общий и детальный анализ (сегментация)	$I(x,y) \rightarrow I_j(x,y) \rightarrow KP_j(x,y)$
3	Выделение и анализ признаков (идентификация)	$KP_j(x,y) \rightarrow \mathbf{K}_{ji} \rightarrow \mathbf{T}_j \rightarrow \mathbf{C}_j$
4	Классификация	$\mathbf{C}_j \rightarrow Q_L$

При разработке информационных технологий для каждого уровня сталкиваются с необходимостью соблюдения ряда требований:

– обеспечение инвариантности к проективным преобразованиям (масштабу, повороту, сдвигу);

– обеспечение устойчивости к шумам и искажениями формы;

– соблюдение жестких временных рамок, т.е. проведение процесса распознавания в реальном масштабе времени;

– обеспечение унификации используемых методов.

Учет этих требований и повышение достоверности получаемой СОИ значимой информации возможен на базе единого информационного подхода и учета моделей представления и обработки ВИ на всех указанных уровнях. Проведенный анализ показал, что существующие основные информационные подходы к моделированию процессов представления и обработки ВИ – структурный (сигнальный), статистический и семантический – не удовлетворяют требованиям практики. Поэтому существует необходимость разработки гибридных моделей обработки ВИ на всех уровнях, которые объединяют разные информационные подходы и позволяют

создавать более эффективные методы и информационные технологии обработки, анализа и распознавания изображений [6].

2. СИГНАЛЬНО-СТАТИСТИЧЕСКАЯ МОДЕЛЬ ИЗОБРАЖЕНИЙ ОБЪЕКТОВ РАСПОЗНАВАНИЯ

На этапе предварительного анализа на основании базы данных и базы знаний, включающих априорную информацию об объекте распознавания, формируются модель изображений объектов распознавания и статистическая модель помеховой ситуации.

Изображение, содержащее объект(ы) на однородном фоне обычно моделируют сечением. Для полутонового (или силуэтного) сечения изображения объекта при моделировании без ограничения общности часто применяется одномерная модель функции интенсивности на участке двустороннего перепада [3]. Модель (ступенчатую функцию) строки изображения объекта I_k получают путем дискретизации и квантования непрерывной модели.

Если полагать, что внутренние геометрические размеры обрабатываемого фрагмента не превышают минимальных размеров ступени, то определить модель фрагмента строки изображения можно следующим образом:

$$I_k = a_0 + a_1 S_k, \quad (1)$$

где a_0 – уровень фона, S_k – бинарный сигнал, a_1 – амплитуда сигнала.

Следует отметить, что преимуществами такой модели является возможность сведения к ней большинства характерных изображений, например, текстурных, и возможность приведения ее к бинарному виду путем присвоения уровню фона нулевых значений, а уровню сигнала – единичных.

Во многих случаях, изображение, полученное с помощью системы формирования изображений, оказывается малопригодным для автоматизированного анализа, из-за низкого качества. Снижение качества изображения может быть вызвано недостаточной четкостью, яркостью или контрастностью, зашумленностью или наличием на нем артефактов [1,2]. К основным причинам появления помех можно отнести следующие:

– внешние факторы – связаны с помехами, формируемыми внешней средой и освещенностью;

– внутренние факторы – обусловлены помехами систем формирования изображений и ошибками на различных этапах преобразования ВИ;

– систематические причины – вызваны неисправностью элементов системы, неправильной настройкой, методическими погрешностями.

Некоторые из этих факторов могут быть частично или полностью скомпенсированы на этапе формирования, но помехи присутствуют практически всегда и их учет и устранение является актуальной задачей при обработке изображений во многих технических приложениях. Зашумленность изображения оказывает большое, а в ряде случаев –определяющее влияние на качество функционирования и эффективность базовых процедур и СОИ в целом. Для обоснованного выбора структуры и параметров информационной технологии предварительной обработки необходимо разработать модель помеховой ситуации. С учетом помеховых факторов статистическую модель обрабатываемого фрагмента строки изображения можно представить следующим образом:

$$I(x, y_i) = \{I_k(x, y_i), R(x, y_i), N(x, y_i), T(x, y_i)\}, \quad (2)$$

где $R(x, y_i)$ – модель неравномерной освещенности объекта; $N(x, y_i)$ – аддитивный гауссовский шум; $T(x, y_i)$ – импульсная помеха.

Данные модели имеют большое значение для решения ряда практически важных задач обработки ВИ.

3. СИГНАЛЬНО-СТАТИСТИЧЕСКАЯ МОДЕЛЬ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ВИ

В автоматизированных системах обработки и распознавания изображений одной из базовых процедур является предварительная обработка. Системы общего назначения работают при отношении сигнал/шум q не менее 10, а проблемно-ориентированные – при q большем 5 [2]. Поэтому снижение уровня помех будет рассматриваться в дальнейшем как цель предварительной обработки (ПрО). Эффективность этой процедуры во многом определяет эффективность СОИ в целом, и, в частности, такой важный параметр как помехоустойчивость. При высоком уровне помех ПрО определяет саму возможность проведения других процедур автоматизированного анализа и распознавания

изображений.

При ПрО и фильтрации сигналов и изображений нашли практическое применение лишь отдельные аспекты математической статистики, основанные на применении крайне упрощенных моделей помех. В частности, при разработке оптимальной согласованной или винеровской фильтрации, основанной на базовом уравнении Винера-Хопфа, полагалось, что помеха является аддитивной и флюктуационной, сам процесс полагался эргодическим, решение искалось в классе линейных систем. Ограничения налагались физической реализуемостью фильтров или вычислительной эффективностью процедуры обработки [6]. При современном экспоненциальном росте возможностей компьютерной техники вычислительная эффективность обработки постепенно отходит на второй план, выдвигая вперед качество обработки. Это позволило на базе анализа основных видов помех и существующих методов предварительной обработки разработать новую методологию синтеза алгоритмов ПрО. Проведенный анализ показал, что все основные методы ПрО (линейная, медианная и гомоморфная фильтрации) несут черты сигнального (взвешивание пикселей в апертуре обработки) и статистического (представление случайным полем и статистическая оценка числовых характеристик) подходов. Это позволило предложить сигнально-статистический подход при моделировании предварительной обработки (табл. 2) [5, 6].

Таблица 2. Сигнально-статистический подход

Вид помехи	Статистический подход к предварительной обработке ВИ	
	Центр группирования	Метод статистической оценки
$R(x, y_i)$	Среднее геометрическое	Среднее в логарифмическом пространстве
$N(x, y_i)$	Математическое ожидание	Среднее в исходном пространстве
$T(x, y_i)$	Медиана	Ранговый метод оценки
$R(x, y_i)$ + $T(x, y_i)$	Среднее геометрическое	Ранговый метод оценки в логарифмическом пространстве
$R(x, y_i)$ + $N(x, y_i)$	Мода	Анализ гистограммы в пространстве вейвлет-преобразования

При таком подходе разработка алгоритмов предварительной обработки ВИ предусматривает три этапа.

На первом этапе осуществляется выбор параметров сигнально-статистической модели:

- исследуются характер и статистические свойства помехи и выбирается адекватный поставленной задаче центр группирования и способ его оценки, отвечающий требованиям состоятельности, несмещенности и эффективности;
- выбирается размерность и форма апертуры (“окна”) $D(n,m)$, в которой производится обработка, где n,m – координаты в апертуре и $\mathbf{D} \subset \mathbf{F}$;
- рассчитываются частотная и импульсная характеристики (функции рассеивания точки), определяется взвешивающая функция, обеспечивающая необходимое преобразование Ф полезного сигнала (изображения) и снижение уровня помех.

На втором этапе реализуется модель предварительной обработки (в скользящем “окне”):

$$I(x,y) = \Phi[\{I'(x+n, y+m)\},\ (n,m) \in D], \quad (3)$$

где $\Phi\{\cdot\}$ – оператор преобразования отсчетов входного сигнала.

Оператор преобразования предусматривает следующую последовательность действий:

– фиксируется обрабатываемый отсчет (пиксель) $I'(x,y)$ и его окрестность $I'(x+n, y+m)$, определяемая размерностью апертуры D;

– в зависимости от координат относительно обрабатываемого отсчета (пикселя) все составляющие окрестности взвешиваются;

– производится оценка выбранного априорно центра группирования взвешенного сигнала (изображения) внутри окрестности;

– значение обрабатываемого пикселя замещается на значение оценки центра группирования $I(x,y)$.

На третьем этапе производится оценка качества функционирования и эффективности предварительной обработки.

Все основные виды предварительной обработки сигналов (изображений) укладываются в схему модели (3). Если учесть возможности использования различных окон и комбинирования разных методов оценки центра

группирования и взвешивания (параметров модели), то очевидны возможности такой модели при решении самых разнообразных задач обработки изображений.

4. СИГНАЛЬНО-СЕМАНТИЧЕСКАЯ МОДЕЛЬ ПРИ КОНТУРНОМ АНАЛИЗЕ ИНФОРМАЦИИ

Обрабатываемое изображение чаще всего является локально неоднородным. Процесс распознавания целесообразно проводить на разных уровнях детализации объекта в зависимости от поставленной задачи. Например, в ряде практически важных задач изображение можно распознать по внешнему контуру объекта (силуэту), в других информативной частью являются мелкие детали объекта. Это соответствует закону перцепции: первоначально выделяется лишь общее, диффузионное представление о предмете, которое затем сменяется более определенным и детальным его восприятием.

Изображение целесообразно представлять в виде последовательности матриц с различным уровнем детальности, что соответствует пирамидальной модели обработки:

$$I(x,y)=\sum_{j=1}^{M} I_j(x,y), \qquad (4)$$

где j – число уровней иерархии на изображении; $I_j(x,y)$ – изображение на j-м уровне иерархии ($j=1,\ldots, M$).

В общем случае изображение $I_j(x,y)$ на каждом уровне есть совокупность изображений отдельных объектов и фона.

В такой постановке задача моделирования процессов обработки ВИ может быть разбита на такие этапы:

– представление изображения в виде структуры изображений “объект – подобъект(ы) – … – подобъект(ы)” в виде (4);

– выделение объектов, соответствующих разным уровням иерархии;

– распознавание объектов и подобъектов.

Существующие методы пирамидального представления информации используют низкочастотную фильтрацию, которая размывает перепады интенсивности и затрудняет выделение контуров. Поэтому целесообразно для представления изображений использовать вейвлет-преобразование, имеющее свойства пространственно-частотной локализации с регулируемой детальностью, подчеркивающее контуры, которое позволит производить обнаружение объектов на верхних уровнях пирамиды с последующей детализацией [2, 5].

Разработанная модель обработки ВИ заключается в следующем:

– на вход модели поступает матричное представление изображения после предварительной обработки;

– матрица интенсивностей подвергается вейвлет-преобразованию с разным уровнем разрешения. Масштабы преобразования выбираются с помощью априорной информации об изображениях в соответствии с целями управления;

– результатом является упорядоченная последовательность матриц с разным уровнем детальности;

– на выбранных уровнях производится выделение контуров и, в случае необходимости, уточнение положения контурного препарата;

– на выходе модели формируется пирамидальное представление контурных препаратов:

$$KP(x,y)=KP_1(x,y)\cup KP_2(x,y)\cup\ldots \\ \ldots\cup KP_M(x,y), \qquad (5)$$

где $KP_j(x,y)$ – контурный препарат на j-м уровне иерархии; M – количество уровней иерархии.

В соответствии с данной моделью обработки энергия изображения концентрируется вблизи наиболее информативной части – перепадов интенсивности, что повышает вероятность правильного распознавания (достижения цели) – семантическую меру информации. Это характерно для семантических методов обработки и представления изображений. С другой стороны, вейвлет-преобразование представимо в виде свертки исходного изображения с вейвлет-функциями разного масштаба, что характерно для сигнального представления. Разработанную информационную модель обработки и представления изображений назовем пирамидальной сигнально-семантической. В качестве базисных функций вейвлет-преобразования могут применяться производные функции Гаусса, гиперболическое вейвлет-преобразование и др. [2, 7].

Информационная технология контурного анализа на базе пирамидальной сигнально-семантической модели позволяет целенаправленно выбирать и изменять количество уровней иерархии, устанавливать пороги обнаружения контурного препарата при изменяющих условиях получения ВИ с целью

повышения достоверности данных для описания геометрической формы объектов. Адаптация порогового уровня при обнаружении контуров происходит в зависимости от цели обработки ВИ, с учетом характера полезного сигнала и помехи в соответствии с заданными критериями эффективности и/или критериальными отношениями, с учетом влияния на другие блоки (процедуры) обработки и систему в целом.

На базе разработанной модели можно реализовать различные подходы к построению иерархических моделей идентификации.

5. ИЕРАРХИЧЕСКИЕ МОДЕЛИ ПРИ СТРУКТУРНО-СТАТИСТИЧЕСКОЙ ИДЕНТИФИКАЦИИ ГЕОМЕТРИЧЕСКОЙ ФОРМЫ ОБЪЕКТОВ

Во многих практических приложениях объект распознавания имеет иерархическую структуру, т.е. может быть представлен в виде уровней: “объект(ы) – подобъект(ы) – элементарный подобъект(ы)”, для которых определено отношение принадлежности: отдельные уровни (подобъекты) располагаются внутри других уровней (объектов), внешних по отношению к ним. Необходимая для работы таких приложений информация заключена в форме объекта, присутствующего на изображении [8].

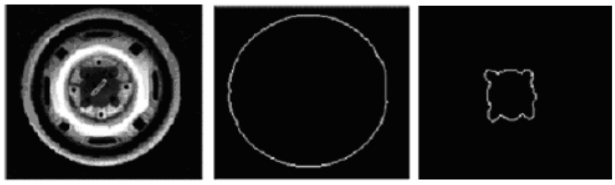
В результате проведения контурного анализа изображений на базе предложенной выше сигнально-семантической модели получают пирамидальное представление геометрических форм объекта и его деталей: набор контурных препаратов. Однако в модели (5) не учитывается цель обработки, определяющая степень детализации, а значит и необходимость получения контурных препаратов, и, следовательно, описаний контуров объектов на разных уровнях детализации. Поэтому предложено ввести параметр $\lambda_{ji} > 0$, определяющий семантическую значимость соответствующего уровня иерархии для получения описания объекта в целом, таким образом, выражение (5) преобразуется к виду

$$KP_{IKO}(x,y) = \bigcup_{j=1}^{M} \lambda_{ji} KP_j(x,y). \qquad (6)$$

Тогда иерархическое контурное описание, может быть представлено выражением

$$\mathbf{K}_{IKO} = \bigcup_{j=1}^{M} \bigcup_{i=1}^{N_j} (\mathbf{K}_{ji})_{\lambda_{ji}}, \qquad (7)$$

где $\mathbf{K}_{ji}$ – упорядоченное множество точек с координатами $\{x_{1iq}^{(j)}, x_{2iq}^{(j)}\}$ контура некоторого *i*-го объекта на плоскости на уровне *j*; *M* – количество уровней; N_j – количество объектов и подобъектов на *j*-м уровне; λ_{ji} – значимость *i*-ого объекта на *j*-ом уровне (от 1 до *N*).



а **б** **в**

Рис. 2 – Пример изображения объекта иерархической структуры (а), контурный препарат первого уровня иерархии (б), контурный препарат j-го уровня иерархии (в).

Поскольку обычно наибольшей информативностью среди точек контура обладают точки наибольшей кривизны (характерные точки), то упорядоченное множество характерных точек *j*-го уровня $\mathbf{T}_j$ может однозначно описывать объект, т.е. справедливо

$$\mathbf{T}_{IKO} = \bigcup_{j=1}^{M} \bigcup_{i=1}^{N_j} (\mathbf{T}_j)_{\lambda_{ji}}. \qquad (8)$$

На базе описаний (7) и (8) могут быть получены топологические признаки объекта (количество и взаимное расположение объектов и подобъектов), спектральные, геометрические и другие характеристики объекта распознавания – вектор контурных (характерных) признаков $\mathbf{C}_{IKO}$, полученный путем преобразования $v_r(\cdot)$ или композиции преобразований $(\ldots(v_2(v_1(\cdot)))$ начального вектора соответствующих признаков. Преобразования $v_r(\cdot)$ осуществляются с учетом значимости уровней иерархии и объектов на них

$$\mathbf{C}_{IKO} = v_m(\ldots(v_2(v_1(\mathbf{K}_{IKO})))),$$
$$\mathbf{C}_{IKO} = v_m(\ldots(v_2(v_1(\mathbf{T}_{IKO})))). \qquad (9)$$

Эффективное моделирование процессов извлечения информации из изображений объектов иерархической структуры предполагает использование разных подходов к комбинированию моделей:

– полученная информация сопоставляется с моделями высокого уровня (начальный уровень *i* соответствует верхнему уровню иерархии), в

случае необходимости эта информация сопоставляется с моделями более низкого уровня, т.е. распознавание проходит по схеме “объект – подобъект” (“сверху вниз” или “нисходящее” описание). Такая схема может использоваться не только при распознавании структурированных объектов, но и при распознавании сцен;

– первоначально находятся подобъекты (начальный уровень *i* соответствует нижнему уровню иерархии), их описания и топологические особенности, затем на основании этих данных выполняется переход на следующий уровень иерархии, т.е. распознавание проходит по схеме “снизу вверх” (“восходящее” описание);

– выполняется одновременное описание объекта на всех уровнях иерархии, вычисление типовых (для данных моделей) признаков и принятие решений по их композиции.

Следует отметить, что использование гибридных моделей предполагает описание изображений объектов на некотором абстрактном уровне. Кроме того, для описания объектов на разных уровнях могут использоваться различные системы признаков. Такие структуры позволяют повысить точность моделирования, упростить и повысить достоверность этапа классификации.

5. ВЫВОДЫ

В статье показано применение информационного подхода к созданию гибридных моделей на разных уровнях обработки изображений. На базе описанных моделей разработано ряд новых эффективных методов, что позволило использовать их для решения широкого круга задач обработки и распознавания изображений.

СПИСОК ЛИТЕРАТУРЫ

[1] П.Г. Катыс, Г.П. Катыс. Системы машинного видения с интеллектуальными видеодатчиками // *Информ. технологии*. – 2001. – № 10. – С. 28-33.

[2] Р. Гонсалес, Р. Вудс. *Цифровая обработка изображений.* – М.: Техносфера, 2005. – 1072 с.

[3] В.Г. Абакумов, С.Г. Антощук, В.Н. Крылов Модели представления и обработки изображений: информационный подход // *Электроника и связь*. Киев. – 2006. – № 5. С. 36-43.

[4] Д. Марр. *Зрение. Информационный подход к изучению представления и переработке зрительных образов*. Пер. с англ. – М.: Радио и связь, 1987. – 400 с.

[5] С.Г. Антощук, В.Н. Крылов. Модели и методы представления и обработки данных при создании информационных технологий // *Оптико-електронні інформаційно-енергетичні технології*. Винница. – 2005. – Вып. 1(9). – С. 16-25.

[6] В.Г. Абакумов, В.Н. Крылов, С.Г. Антощук. Предварительная обработка сигналов и изображений // *Электроника и связь.* Киев. – 2004. – № 21. – С. 64-67.

[7] С.Г. Антощук, В.Н. Крылов. Обработка изображений в области гиперболического вейвлет-преобразования // *Автоматика. Автоматизация. Электротехнические комплексы и системы*. Херсон. – 2003. – № 2. – С. 7-10.

[8] С.Г. Антощук, А.А. Николенко, О.Ю. Бабилунга. Модель структурированного объекта с учетом семантической значимости // *Международная научно-техническая конференция “Искусственный интеллект. Интеллектуальные системы” (ИИ-2008).* Донецк, 2008. – С. 42-45.

Антощук Светлана Григорьевна, *Директор института компьютерных систем Одесского национального политехнического университета, заведующая кафедрой информационных систем, доктор технических наук, профессор.*

Научные интересы: информационные управляющие системы и технологии, компьютерные системы искусственного интеллекта.

Бабилунга Оксана Юрьевна, *Кандидат технических наук, доцент кафедры информационных систем Одесского национального политехнического университета.*

Научные интересы: системы и средства искусственного интеллекта, обработка и распознавание изображений.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

HYBRID INFORMATION MODELS IN IMAGE PROCESSING SYSTEMS

S. Antoshchuk, O. Babilunga

Department of the Information Systems,
Institute of the Computer Systems,
Odessa National Polytechnic University,
Shevchenko prospect, 1, Odessa, 65044, Ukraine,
svetlana_onpu@mail.ru, babilunga@mail.ru

Abstract: *In this article the approach to the development of hybrid information models in image processing systems is presented. This approach allowed creating more effective methods and information technologies of the image processing, analysis and recognition.*

Keywords: *information technologies, visual information, image processing and recognition, hybrid models, contour analysis.*

INTRODUCTION

Visual information processing systems (VIPS) are often used in the modern information and controlling systems (ICS). These systems are used for the solution of the most various tasks: measuring and control of linear and angular sizes of stationary and moving objects, number of objects and count of products, control of goods shape and determination of deviations from standard shapes etc. [1, 2].

At the present the domain of VIPS application is limited by their insufficient efficiency and universality. Therefore the quality of processing of visual information, presented as images, must increase. This is an important scientific and practical problem, which must be solved by the effective modeling of VIPS and analysis of processes, which take place in the systems [3, 4].

1. ANALYSIS OF INFORMATION PROCESSES IN IMAGE PROCESSING SYSTEMS

VIPS are the complex systems, working in the conditions of non-determination of initial information. Their basic function is the extraction of essential information, determining efficiency of ICS work on the whole. For their design it is expedient to use system approach. The hierarchy of VIPS application aims is reflected in a system model (fig.1).

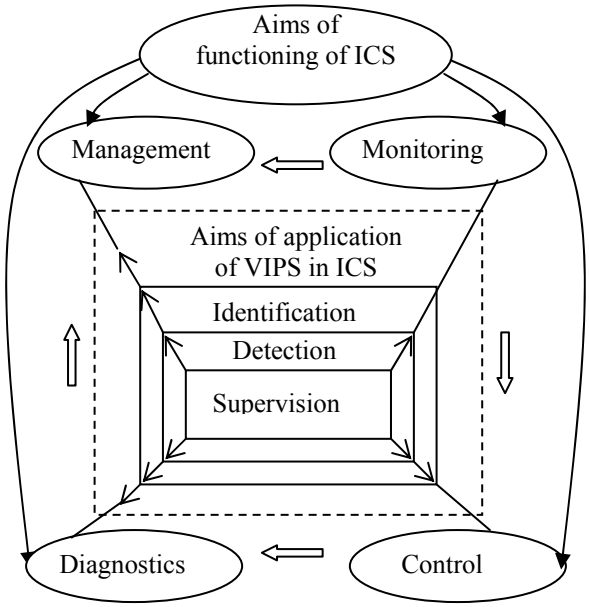


Fig. 1 – The system model of VIPS application in ICS

In accordance with the system model, the obtained visual information in ICS can be designated for supervision an object (collection, transformation and preview processing of information, visualization), detecting an object or measuring of information about it (control, determination of deviations) and identification of an object or its characteristic features (diagnostics, estimation of management object parameters and classification conclusion) [5].

Existent VIPS include the image generation systems and image processing systems. The image

generation systems carry out the perception and registration of visual information on the carrier of data. The choice of the concrete type of visual information reading is produced on the basis of the known (exactly or approximately) features of the researched objects taking into account the necessary operating requirements.

The analysis of the basic information approach to the design of visual information presentation and processing in image processing systems, in regarding to the system model, allows selecting four levels of visual information processing: preview processing, general and detailed analysis (segmentation), selection and analysis of features (identification), classification.

In accordance with the proposed system model, the role of image processing systems can be different. In creating them it is expedient to apply information approach. Three closely connected problems consider in terms of this approach: entrance presentation, processing and output presentation of images.

2. DEVELOPMENT OF HYBRID INFORMATION MODELS

The conducted analysis showed that the existing basic information approaches to the design of visual information presentation and processing – structural (signal), statistical and semantic – dissatisfied the requirements of practice. Therefore there is a necessity of development of hybrid models of visual information processing on all the levels, which unite different informative approaches and allow creating more effective methods and information technologies of images processing, analysis and recognition.

The next information models are offered and analyzed in the article [5, 6, 7, 8]:

– signal-statistical model of image object recognition;

– signal-statistical model of visual information preview processing;

– signal-semantic model at the contour analysis of information;

– hierarchical models in terms of the structural-statistical identification of geometrical object shapes.

It should be noted that the use of the hybrid models supposes the description of object images at some abstract level. In addition, various systems of features can be used in order to describe objects at different levels. Such structures allow to improve exactness of modeling and to increase reliability of the classification.

4. CONCLUSION

The article presents application of the information approach to creating of hybrid models at different levels of image processing. On the base of the described models a row of new effective methods is developed. These models and methods may be applied for the solution of wide circle of tasks of image processing and recognition.

5. REFERENCES

[1] P.G. Katys, G.P. Katys. Systems of Machine Vision with Intelligent Videosensors // *Inform. Technologies.* – 2001. – No. 10. – Pp. 28-33. (in Russian).

[2] R. Gonsales, R. Woods. *Digital Image Processing.* – Moscow: Technosphere, 2005. – 1072 p. (in Russian).

[3] V.G. Abakumov, S.G. Antoshchuk, V.N. Krylov. Models of Image Presentation and Processing: Information Approach // *Electronics and Connection.* – 2006. No. 5. – Pp. 36-43. (in Russian).

[4] D. Marr. *Vision. The Information Approach to the Learning of Visual Pattern Presentation and Processing.* – Moscow: Radio and Connection, 1987. – 400 p. (in Russian).

[5] S.G. Antoshchuk, V.N. Krylov. Models and Methods of Data Presentation and Processing at Creation of Information Technologies // *Optiko-Electronic Informative-Power Technologies*. – 2005. – Vol. 1(9). – Pp. 16-25. (in Russian).

[6] V.G. Abakumov, V.N. Krylov, S.G. Antoshchuk. Preview Processing of Signals and Images. *Electronics and Connection*. – 2004. – No. 21. – Pp. 64-67. (in Russian).

[7] S.G. Antoshchuk, V.N. Krylov. Image Processing in Area of Hyperbolic Wavelet-Transformation. *Automatization. Automation. Electrical Engineerings Complexes and Systems.* – 2003. – No. 2. – Pp. 7-10. (in Russian).

[8] S.G. Antoshchuk, A.A. Nikolenko, O.Yu. Babilunga. Model of the Structure Object Regarding the Semantic Meaning // *International Scientific and Technical Conference is "Artificial Intelligence. Intellectual Systems" (II-2008)*. Donetsk. – 2008. – Pp. 42-45. (in Russian).

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

A MODIFIED A* ALGORITHM FOR ALLOCATING TASK IN HETEROGENEOUS DISTIRBUTED COMPUTING SYSTEMS

Nirmeen A. Bahnasawy [1], Gamal M. Attiya [2], Mervat Mosa [1], Magdy A. Koutb [2]

[1] Dept. of Computer Science and Engineering, Faculty of Engineering, Menoufia University
nirmeen_a_wahab@hotmail.com, nirmeen23875@yahoo.com
[2] Dept. of Automatic Control Engineering, Faculty of Engineering, Menoufia University

Abstract: *Distributed computing can be used to solve large scale scientific and engineering problems. A parallel application could be divided into a number of tasks and executed concurrently on different computers in the system. This paper provides an optimal task assignment algorithm under memory constraints to minimize required time of finishing a parallel application. The proposed algorithm is based on the optimal assignment sequential search (OASS) of the A* algorithm with additional modifications. This modified algorithm yields optimal solution, lower time complexity, reduces the turnaround time of the application and considerably faster compared with the sequential search algorithm.*

Keywords: *parallel processing, task assignment, distributed computers, heterogeneous processors.*

1. INTRODUCTION

A distributed computing system is characterized by a topology, availability of computers and communication resources. This system consists of heterogeneous computers interconnected through a communication network. Each computer has computation facilities, its own memory, communication capacity and propagation delay between two computers [5]. Such a system can be employed to solve large scale scientific and engineering problems, where, an application could be divided into a number of tasks and executed concurrently on different computers in the system. The distribution of tasks onto different computers of the system is called *allocation problem*. Such a problem is known to be NP-hard, except in a few special cases with strict assumptions.

Several approaches were by many researchers to solve the allocation problem. In [3,8], approaches are developed for allocating and scheduling tasks of a parallel application among computing sites of distributed computing system to achieve some objectives under defined constraints. Shen and Tsia [18] first used the A*algorithm for the task-assignment problem. They ordered the tasks considered for assignment simply starting with task 1 at the tree's first level, task 2 at the second and so on. Kafel [10] considered the assignment problem to minimize the parallel program completion time, they proposed algorithm based on A* algorithm called "Optimal Assignment Sequential Search" (OASS). Also an optimal static scheduling of an arbitrary structured task graph to an arbitrary number of homogeneous processors is discussed on the A*search technique with a computationally efficient cost function and number of state-space pruning technique [16]. Note that, A* is a best-first search algorithm, which has been used extensively in artificial intelligence problem solving. Programmers can use the algorithm to search a tree or a graph.

The problem of minimizing the total cost subject to both memory and processing capacity constraint is discussed in [14] by taking into account the limited capacities of the resources that constitute the communication network (LANs, WANs, and direct link). Task allocation problem is discussed also in [1] where, A* algorithm (A*RS) can be implemented to improve the performance of the earlier algorithm then allocate the multiple tasks that requires more time and space for the solution.

In [5], a simulated annealing technique is developed to quickly find a near optimal solution to make the system more reliable in inter processor communication times under conditions imposed by both the application and system resources. In [12], a genetic algorithm (GA) is developed to finding approximate solutions for problems with very large decision spaces by applying it on the task matching problem independent task. In [16, 17], a new hybrid genetic algorithm is used to solve the task scheduling problem in heterogeneous computing

system which is guarantee that every feasible schedule is reachable with some probability. Many algorithms are compared, evaluated, and get analyzed to present the scheduling task problem [7]. Heuristic algorithm [14] showed that the most effective non-evolutionary known method for scheduling independent tasks in heterogeneous environment is the min-min heuristic.

In this paper, the main objective is to study task assignment problem in distributed systems comprising networked heterogeneous computers and then develop a new technique for obtaining optimal solution to the given problem to minimize the turn around time of the application A comparison is done between the (OASS) and the modified algorithm in assigning a given tasks to a network of processors to minimize the required time for program completion.

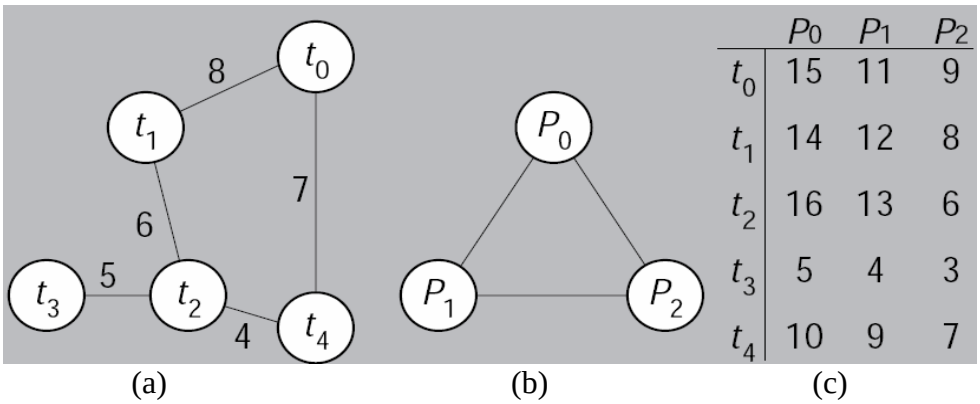
Note that, minimizing the turn around time of a parallel application can be achieved only when the work load is balancing distributed on the different processors of the system

The rest of this paper is organized as follow. Section 2 defines the task assignment problem. Section 3 describes task assignment using optimal sequential search of A* algorithm. Section 4 presents the modified algorithm. Section 5 presents the simulation result, and finally section 6 presents the conclusions.

2. PROBLEM DEFINITION

Task allocation problem may be stated informally as the problem of allocating tasks of a parallel application onto computers of distributed computing system to optimize some performance measures as finishing time [4,6].

In solving the allocation problem, we use the *task interacting graph* model, in which the parallel program is represented by an undirected graph: $G_T = (V_T, E_T)$, where V_T is the set of vertices, $\{t_1, t_2, ..., t_m\}$, and E_T is a set of edges labeled by the communication requirements among tasks. We can also represent the network of processors as an undirected graph, where vertices represent the processors, and the edges represent the processors' communication links. We represent the interconnection network of n processors, $\{p_1, p_2, ..., p_n\}$, by an $n \times n$ link matrix **L**, where an entry $\mathbf{L}_{ij}$ is 1, if processors i and j are directly connected, and 0 otherwise. We do not consider the case where i and j are not directly connected. We can execute a task t_1 from the set V_T on any one of the system's n processors. Each task has an associated execution cost on a given processor. A matrix **X** gives task execution costs, where $\mathbf{X}_{ip}$ is the execution cost of task i on processor p. Two tasks, t_i and t_j, executing on two different processors, incur a communications cost when they need to exchange data. Task mapping will assign two communicating tasks to the same processor or to two different, directly connected processors. A matrix **C** represents communication among tasks, where $\mathbf{C}_{ij}$ is the communication cost between tasks i and j, if they reside on two different processors.



	P_0	P_1	P_2
t_0	15	11	9
t_1	14	12	8
t_2	16	13	6
t_3	5	4	3
t_4	10	9	7

(a) (b) (c)

Fig. 1 – An Example (a) a task graph (b) processor Network (c) Execution cost matrix

A processor's load comprises all the execution and communication costs associated with its assigned tasks. The time needed by the heaviest-loaded processor will determine the entire program's completion time. The task-assignment problem must find a mapping of the set of m tasks to n processors that will minimize program completion time. Task mapping, or assignment to processors, is given by a matrix **A**, where $\mathbf{A}_{ip}$ is 1, if task i is assigned to processor p, and 0 otherwise. The following equation then gives the load on p:

$$Z = \sum_{i=1}^{m} X_{ij} \bullet A_{ip} + \sum_{\substack{q=1 \\ (q \neq p)}}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} (C_{ij} A_{ip} A_{jq} L_{pq})$$

The first part of the equation is the total execution cost of the tasks assigned to p. The second part is the communication overhead on p. A_{ip} and A_{jq} indicate that task i and j are assigned to two different processors (p and q), and L_{pq} indicates that p and q are directly connected. To find the processor with the heaviest load, you need to compute the load on each of the n processors. The optimal assignment out of all possible assignments will allot the minimum load to the heaviest-loaded processor.

3. SEQUENTIAL SEARCH ALGORITHMS (OASS)

Task assignment using the A* algorithm will be occurred as follows. For a tree search, it starts from the root, called the start node (usually a null solution of the problem). Intermediate tree nodes represent the partial solutions, and leaf nodes represent the complete solutions or goals. A cost function f computes each node's associated cost. The value of f for a node n, which is the estimated cost of the cheapest solution through n, is computed as: $f(n) = g(n)+ b(n)$, where $g(n)$ is the search-path cost from the start node to the current node n and $b(n)$ is a lower-bound estimate of the path cost from n to the goal node (solution). To expand a node means to generate all of its successors or children and to compute the f value for each of them.

The nodes are ordered for search according to cost; that is, the algorithm first selects the node with the minimum expansion cost. The algorithm maintains a sorted list, called OPEN, of nodes (according to their f values) and always selects a node with the best expansion cost. Because the algorithm always selects the best-cost node, it guarantees an optimal solution.

For the task-assignment problem under consideration,

- The search space is a tree;
- The initial node (the root) is a null assignment node—that is, no task is assigned;
- Intermediate nodes are partial-assignment nodes—that is, only some tasks are assigned; and
- A solution (goal) node is a complete assignment node—that is, all the tasks are assigned.

```
Generate a random solution
Let S_Opt be the cost of this solution
Build the initial node s and insert it into the list OPEN
Set f(s) = 0
Repeat
    Select the node n with smallest f value.
     Repeat
     Make memory test step
      If (n is true)
    if (n is not a Solution )
       Generate successors of n
        for each successor node n' do
            if (n' is not at the last level in the search tree)
                  f(n') = g(n') + b(n')
           else   f(n') = g(n')
           if (f(n') <= S_Opt)
                 Insert n' into OPEN
         endif
       end for
      else select (n+1)
    end if
    if (n is a Solution)
      Report the Solution and stop
Until (n is a Solution) or (OPEN is empty)
```

Fig. 2 – The Optimal Assignment with Sequential Search algorithm (OASS)

To compute the cost function, $g(n)$ is the cost of partial assignment *(A)* at node n the load on the heaviest-loaded (p); this can be done using the equation from the previous section. For the computation of $b(n)$, two sets T_p (the set of tasks that are assigned to the heaviest-loaded p) and U (the set of tasks that are unassigned at this stage of the search and have one or more communication link with any task in set T_p) are defined. Each task t_i in U will be assigned either to p or any other processor q that has a direct communication link with p. So, associate two kinds of costs with each t_i's assignment: either X_{ip} (the execution cost of t_i on p) or the sum of the communication costs of all the tasks in set T_p that have a link with t_i. This implies that to consider t_i's assignment, we must decide whether t_i should go to p or not (by taking these two cases' minimum cost). Let cost (t_i) be the minimum of these two costs, then we compute $b(n)$ as:

$$b(n) = \sum_{t_i \in U} \cos t(t_i)$$

The A* algorithm for the task-assignment problem has been used early. The tasks are ordered considered for assignment simply by starting with task 1 at the tree's first level, task 2 at the second, and so on.

OASS algorithm [10] uses the A* search technique, but with two distinct features. First, it generates a random solution and prunes all the nodes with costs higher than this solution during the optimal-solution search see Figure 2. This is because the optimal solution cost will never be higher than this random-solution cost. Pruning unnecessary nodes not only saves memory, but also saves the time required to insert the nodes into OPEN. Second, the algorithm sets the value of *f(n)* equal to *g(n)* for all leaf nodes, because for a leaf node n, *b(n)* is equal to 0. This avoids the unnecessary computation of *b(n)* at the leaf nodes.

4. THE MODIFIED ALGORITHM

In this section a new task assignment algorithm is presented. The basic idea is to choose the task to be assigned at each level. The assignment problem is represented as A*algorithm for traversing the tree nodes searching for an optimal solution.

The proposed algorithm handles tasks at the tree levels according to the task of higher connectivity, i.e., the task with the largest number of neighbors and then test the memory constrains, This constrain shows; for an assignment A, the total memory required by all tasks assigned to a processor p must be less than or equal to the available memory capacity of the processor p. Let m_i denotes the amount of memory required for processing a task i and M_p defines the available memory capacity at the processor p, then the following inequality must hold at each processor p in the system:

$$\sum_i m_i A_{ip} \le M_p \ \forall p$$

i.e., before distribution, the program compares memory capacities of the chosen task to be run on the chosen processor ; if their memory capacities are equal or memory capacity of task is less than the memory capacity of processor or memory capacity of processor permits to run that task, the program will complete, and if the memory capacity of task is larger than the memory capacity of processor, the program will search for another one so, the chosen task will choose the next processor to make the same test and neglect that one which is not fit then generate node, and so on until the program complete; this step is called memory constrain step (1) as illustrated in Figure 3.

The algorithm starts by reordering the application tasks according to the task of more connectivity, i.e., calculate the sum of the communication lines for all tasks then choose the highest one, if two tasks have the same connectivity degree; it considers the task of higher communication requirements, then it generates a random solution and prunes all the nodes with costs higher than this solution during the optimal-solution search, this is because the optimal solution cost will never be higher than this random-solution cost. Pruning unnecessary nodes not only saves memory, but also saves the time required to insert the nodes into OPEN, also the algorithm sets the value of *f(n)* equal to *g(n)* for all leaf nodes, because for a leaf node n, *b(n)* is equal to 0, this avoids the unnecessary computation of *b(n)* at the leaf nodes.

Our proposed algorithm is effective in reducing the number of nodes generated (and that are expanded) without sacrificing the optimality criteria. This property allows to reduce mathematical computations such as computing *f(n)* for latest nodes.

The proposed algorithm will be applied on the previous example to study in details the comparison between it and the sequential search algorithm.

Consider (Figure 1), by applying the idea of the proposed algorithm in (Figure 3). The new order of tasks list is obtained as follow: $\{t_2,t_0,t_1,t_4,t_3\}$, the number of resulting node expansion is reduced compared with sequential search algorithm.

```
Generate a random solution
 Let S_Opt be the cost of this solution
For (i=0;i<=n;++)
    Compute connectivity degree and communication requirements.
    If (ti connectivity degree>ti+1 connectivity degree)
       Set ti in ordered task list
          If (ti connectivity degree= ti+1 connectivity degree)
            Compute the communication requirements.
                If (ti communication requirements>ti+1communication requirements)
                    Set ti in ordered task list
                Else
                    If (ti communication requirements=ti+1 communication requirements)
                     Set ti in ordered task list
          Else
              Set ti+1in ordered task list
    Else
       Set ti+1in ordered task list
End for
 Build the initial node s and insert it into the list OPEN
 Set f(s) = 0
 Repeat
     Select the node n with smallest f value.
    Make memory test step
     If(n is true)
     if (n is not a Solution )
        Generate successors of n
         for each successor node n' do
              if (n' is not at the last level in the search tree)
                    f(n') = g(n') + b(n')
             else   f(n') = g(n')
             if (f(n') <= S_Opt)
                    Insert n' into OPEN
```

```
    Else select n+1
       end for
     end if
     if (n is a Solution)
       Report the Solution and stop
 Until (n is a Solution) or (OPEN is empty)
```

Fig. 3 – The Proposed Algorithm

Consider t_2 with three communication links, t_0 which has two communication links like: t_0, t_1, t_4, then the highest communication requirements is chosen, and so on. We get 11 nodes are generated by applying the modified algorithm When OASS algorithm is applied on the same example, 14 nodes are generated compared with these results yield the same optimal solution. This reduction in results saves memory, increase speedup and the performance of program. The modified algorithm has considerably fewer memory requirements than OASS algorithm.

5. SIMULATION REULTS

To test the performance of the algorithm first, simulated by C# ver.5.1and then, a simulation environment in acer core due processor with 1.73 speedup.

To test the modified algorithm the data collect for task graphs of 2 to 30 nodes and processor graphs of 2, 4, 8 nodes. Figure 4 shows node generated of A*O, OASS and modified algorithm on 4 processor nodes. As a result, the number of generating nodes decreases, the running time of program decreases, so the system required memory decreases and then memory efficiency increases in Malg.

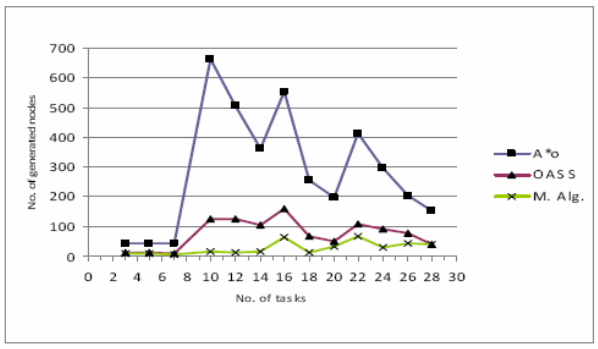


Fig. 4 – Generated nodes case of 4 processors

We also obtain a comparison between the same three algorithms in the running time parameter. The modified algorithm shows a substantial improvement over (OASS) algorithm. Figure 5 shows collecting data for task graph of 3 to 28 tasks and processor graph of four nodes, the results of running times in seconds are obtained, when running time decreased, the speedup of program increased so, the modified algorithm behaves better performance than (OASS) algorithm.

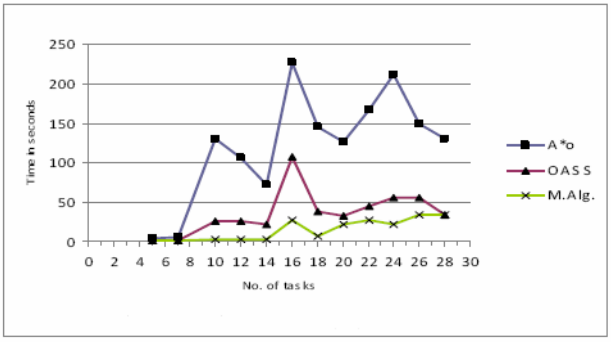


Fig. 5 – Runing times using 4processors

Figures 6 (a, b, c) show generated nodes of A*O, OASS and modified algorithm on (2, 4, 8) processor nodes respectively, the modified algorithm shows a substantial improvement over (OASS) algorithm and A*O algorithm.

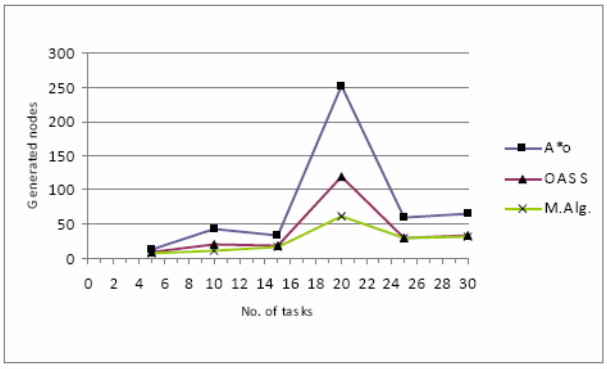


Fig. 6(a) – Generated nodes caseof 2 processors

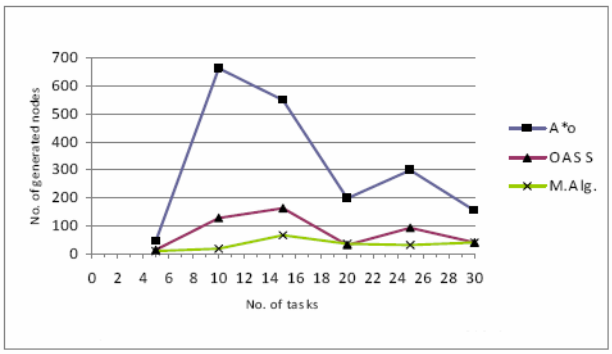


Fig. 6(b) – Generated nodes case of 4 processors

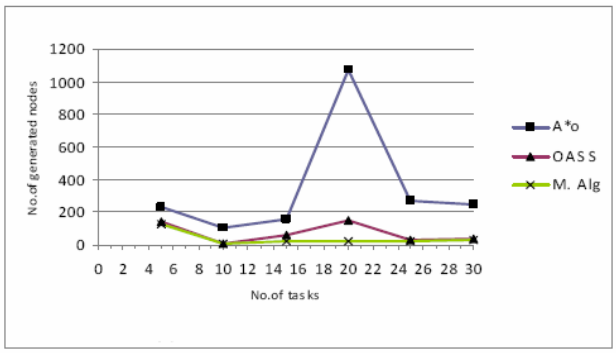


Fig. 6(c) – Generated nodes case of 8 processors

In addition, the improvement in the modified algorithm performance, it efficiently uses the system memory as shown in Figure (7).As the total number of nodes decreases, the required memory saving decreases. Note that, the saving rate is defined as:

$$(1-(G_{(Malg.or\ OASSalg)} / G_{A^*alg})),$$

Where, $G_{(M.alg.or\ OASS\ alg)}$ is the actually number of nodes generated by the two algorithms with respect to A* algorithm.

Figures 7 (a, b, c) show the memory saving rates on (2, 4, 8) processor nodes respective

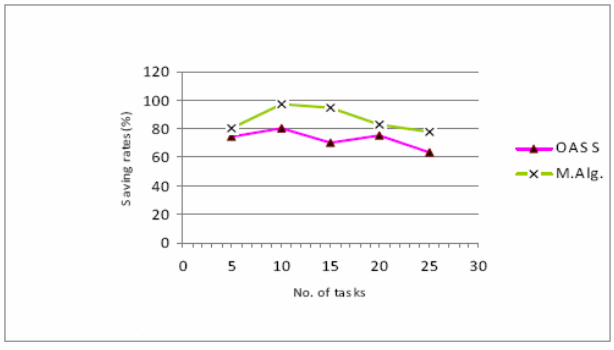


Fig. 7(a) – Memoryrates case of 2 processors

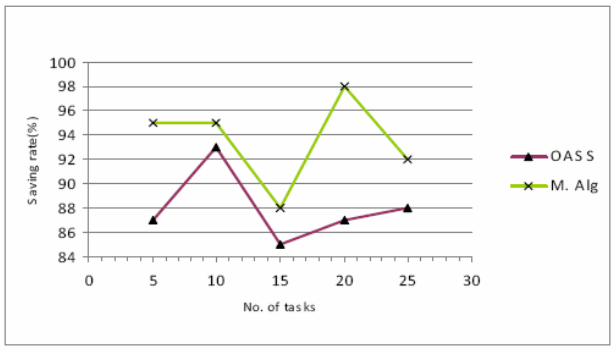


Fig. 7(b) – Memory rates case of 4processors

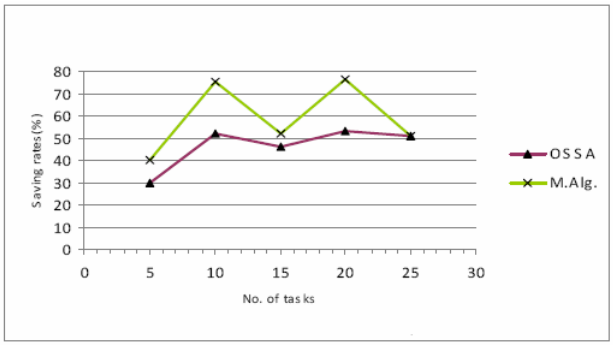


Fig. 7(c) – Memory rates case of 8 processors

Figure (8) presents simulation results of node generated from the previous three algorithms on 4 processor nodes taking into account the memory constrains, if it is included or excluded.

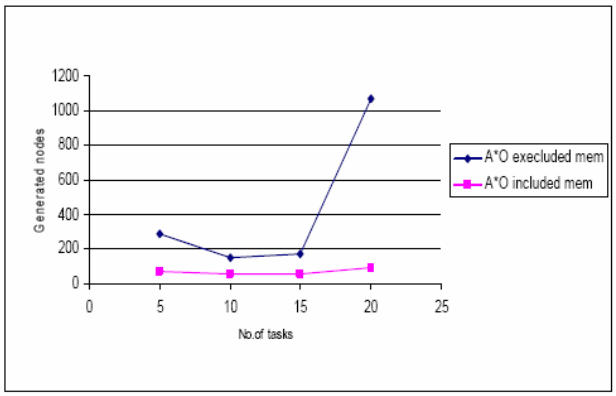


Fig. 8(a) – Simulation results of containing memory conditions of A*O

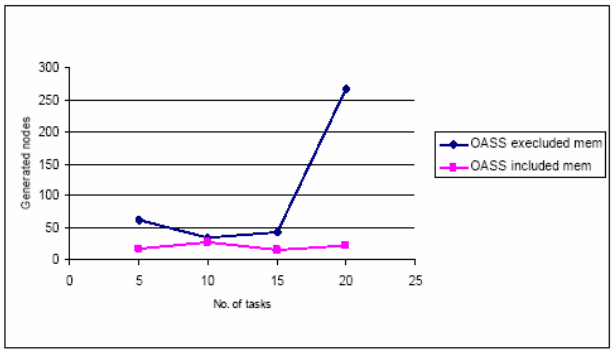


Fig. 8(b) – Simulation results of containing memory conditions of OASS

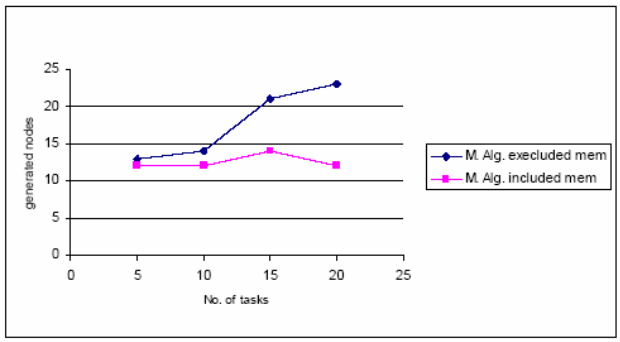


Fig. 8(c) – Simulation results of containing memory conditions of M.Alg.

6. CONCLUSIONS

In this paper, the task allocation problem is studied and a new algorithm for allocating application graphs on to a system of heterogeneous processors is presented. The performance of the proposed algorithm has been investigated in terms of memory efficiency, running time, and saving memory. This has been carried out by using a set of randomly generated application graphs under different conditions like; the communication data rates, the capacities memory of various processors and tasks, and the weight cost of running tasks.

Based on the experimental study, the simulation results show, the proposed algorithm is superior in terms of efficiency and quality in using the memory and also speedup compared with (OASS) which are most important performance measures of evaluating a parallel computer system.

7. REFERENCES

[1] A. K. Tripathi, B. K. Sarker, N. Kumar and D. P. Vidyanhi, "*Multiple Task Allocation with Load Considerations*", Int. Journal of Information and Computing Science (IJICS), Vo1. 3, No, pp.3634, 2000.

[2] Gamal Attiya and Yskandar Hamam, "*Optimal Allocation of Tasks onto Networked Heterogeneous Computers Using Minimax Criterion*", Proceedings of International Network Optimization Conference (INOC,03), pp. 25-30, Evry/Paris, France, 2003.

[3] Gamal Attiya and Yskandar Hamam, "*Hybrid Algorithm for Mapping Parallel Applications in Distributed Systems*", Fifth International Conference on PRAM, Poland, 7-10 September 2003.

[4] Gamal Attiya and Yskandar Hamam, "*Performance Oriented Allocation in Heterogeneous Distributed Systems*", European Simulation and Modeling (ESM 2003) Conference, University of Naples II, Naples, Italy, pp. 27-29 October 2003.

[5] Gamal Attiya and Yskandar Hamam, "*Task Allocation for Maximizing Reliability of Distributed Systems: A simulated Annealing Approach*", J. Parallel Distributed Computer, 66, pp. 1259-1266, 2006.

[6] Haluk Topcuoglu, Salim Hariri, and Min-You Wu, "*Performance- Effective and Low-Complexity Task Scheduling for Heterogeneous Computing*", (IEEE Transactions on Distributed Systems, vol. 13. no. 3 march 2002.

[7] I. Ahmed, Y Kwak, and M.-Y. Wu, "Analysis, *Evaluation, and Comparison of Algorithms for Scheduling Task Graphs on Parallel Processors*", 2nd Int'l Symposium on Parallel architectures, Algorithms, and Networks, I-SPAN. Beijing. China. Proc. Of the 1996.

[8] Kamer Kaya a, Bora Ucar a, Cevdet Aykanat, Murat Ikinci, "*Task assignment in heterogeneous computing systems*" J. Parallel Distributed Computers. 66, pp. 32 – 46, (2006).

[9] L. W. Dowdy, E.Rosti, E. Smirni, G. Serazzi, and K. C. Sevcik, "*Processor Saving Scheduling Policies for Multiprocessor Systems*" IEEE Trans. Comput., vol. 47, no. 2, Feb 1998.

[10] M. Kafil, "A*n Optimal Task Assignment Algorithms in Distributed Computing Systems*", IEEE Concurrency, 98, pp 42-49, September 1998.

[11] M. Mezmaz, N. Melab, and E.-G. Talbi, "*An Efficient Load Balancing Strategy for Grid-Based Branch and Bound Algorithm*", Parallel Computing 33, pp. 302-313. February 2007.

[12] S. Woo, S. Yang, S. Kim, and T. Han, "*Task Scheduling in Distributed computing Systems with a Genetic Algorithm*", IEEE, New York, 1997.

[13] Sih, E.A.Lee, "*A Compile-Time Scheduling Heuristic for Interconnection-Constrained Heterogeneous Processor Architectures*", IEEE Trans. Parallel Distributed Systems4, pp. 175-186, February 1993.

[14] W. Boyer, G. Hura, "*A New Min-Span Heuristic Algorithm for Task Scheduling in Heterogeneous Systems*" Proceedings of the Sixth Biennial World conference On Integrated Design and Process Technology, 23. pp. 69, June 2002.

[15] Yskandar Hamam, Khalil S. Hindi, "*Assignment of Program Modules to Processors: A Simulated Annealing Approach*", European Journal of Operational Research 122, pp. 509-513. 2000.

[16] Yu-Kwong Kwok, Ishfaq Ahmad, "*On Multiprocessor Task Scheduling Using Efficient State Space Search Approaches*", J. Parallel Distributed Computing 65, pp. 1515-1532, (2005).

[17] Yi-Wen Zhong, Jian-Gang Yang, and Heng-Nian, "*A Hybrid Genetic Algorithm for Tasks Scheduling in Heterogeneous Computing Systems*" Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, PP. 26-29, August 2004.

[18] Z.-C. Shen and W.-H. Tsai, "*A Graph Matching Approach to Optimal Task Assignment in Distributed Computing System Using a Minimax Criterion*", IEEE Trans. Computers, Vol. C-34, No. 3, Mar., pp. 197–203 1985.

Nirmeen A. Wahab Al Bahnasawy *is a PhD candidate in the Dept. of Computer Science and Engineering, Faculty of electronic Engineering, Menoufia University, Egypt. Her research interests include parallel processing, distributed computing, task allocation and scheduling.*

Mervat Mosa *is PhD in the Dept. of Computer Science and Engineering, Faculty of electronic Engineering, Menoufia University, Egypt. Her research interests include parallel processing and optical computing.*

Gamal M. ATTIYA *graduated in 1993 and obtained his MSc degree in computer science and engineering from the Menufiya University, Egypt, in 1999. He received PhD degree in computer engineering from the University of Marne-La-Vallée, Paris-France, in 2004. He is currently Lecturer at the department of Computer Science and Engineering, Faculty of Electronic Engineering, Menufiya University, Egypt. His main research interests include distributed computing, task allocation and scheduling, computer networks and protocols, congestion control, QoS, and multimedia networking.*

Magdy A. Koutb *received the Eng. Degree from the university of Menofiya, Egypt, in 1977 and M.Sc.degree in 1981 from the university of AlAzhar, Egypt, and the Ph.D. degree from the university of Silesia, Poland in 1985. He has been a Professor since 1997 at the Faculty of Electronic Eng., Menoufiya university. In 2003 he was appointed as Vice-Dean of Post-Graduate Studies and Cultural affairs of the same faculty. He has extensive experience in Computer Engineering and Industrial Electronics. His main research interests include distributed systems, network security and intelligent control systems.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

ВИБІР НАЙКОРОТШОГО З АЛЬТЕРНАТИВНИХ СТИСНУТИХ БЛОКІВ ДИНАМІЧНИХ КОДІВ ХАФМАНА У ФОРМАТІ PNG

Олександр Шпортько

Кафедра інформатики та прикладної математики
Рівненський державний гуманітарний університет, Україна
33028, м. Рівне, вул. С. Бандери, 12
E-mail: chportko@ukr.net, chportko@yandex.ru

Резюме: *у статті пропонується алгоритм генерування альтернативних стиснутих блоків динамічних кодів Хафмана для кожного блоку даних, вибору найкоротшого стиснутого блоку з альтернативних та ітеративного зменшення його розміру для покращення компресії зображень у форматі PNG. Детально розглядаються способи оцінки розміру блоку кодів Хафмана за абсолютними частотами його елементів. Наведено фрагмент підпрограми мовою C, яка дозволяє точно визначати розмір стиснутого блоку за допомогою принципу генерації динамічних кодів Хафмана та окремого опрацювання малих частот елементів, витрачаючи для цього в середньому не більше часу, ніж для наближеної оцінки розміру з використанням ентропії. Як показують експерименти, реалізація запропонованого алгоритму дозволяє покращити показники стиснення переважної більшості зображень у форматі PNG на 2 – 6 %.*

Ключові слова: *безвтратне стиснення зображень, динамічні коди Хафмана, формат графічних файлів PNG.*

ВСТУП

Формат графічних файлів PNG був створений 1 жовтня 1996 року для ефективного збереження растрових зображень без втрат після того, як компанія Unisys почала вимагати плату за використання формату GIF [1]. На сьогодні дизайнери та розробники Web-сайтів найчастіше зберігають фотореалістичні зображення у форматі JPEG, а дискретно-тонові – у форматі PNG. Крім цього, формат PNG найчастіше використовується для зберігання зображень, де втрати неприпустимі (наприклад, для рентгенівських знімків). Саме тому цей формат підтримується більшістю сучасних програм для перегляду і створення зображень. Проблема підвищення ефективності стиснення зображень у форматі PNG є актуальною сьогодні і буде актуальною в найближчому майбутньому, оскільки навіть наближення до її розв'язання дозволяє зменшити розміри відповідних файлів, що, в свою чергу, дає змогу прискорити їх завантаження з мережі та підвищити ефективність використання дискового простору. В даній роботі детально описується спосіб покращення показників стиснення зображень у цьому форматі за рахунок генерування та порівняння розмірів альтернативних стиснутих блоків, розгляд яких започатковано в [2].

1. ПРИНЦИПИ СТИСНЕННЯ ЗОБРАЖЕНЬ У ФОРМАТІ PNG. ПОСТАНОВКА ЗАДАЧІ

Будь-яке стиснення даних можливе за рахунок зменшення надлишковостей. Чим більше видів надлишковостей виявляє і опрацьовує компресор і чим краще він ці надлишковості усуває – тим якісніше він зможе стиснути націлені на таку обробку дані. Формат PNG орієнтований на зменшення надлишковостей чотирьох типів:

- між сусідніми пікселами, що мають, як правило, близькі кольори;
- між однаковими фрагментами зображення;
- між однаковими змінами кольорів;
- між переважаючими кольорами пікселів зображення.

Пікселі зображення записуються у PNG-файли найчастіше по рядках зверху вниз, а у кожному рядку – послідовно зліва направо (як символи у текстах), формуючи тим самим вхідний потік. Перед кодуванням ці пікселі можуть бути попередньо оброблені предикторами, принцип дії яких розглядається наприкінці цього розділу. У PNG-файлах, що

використовуються на сьогодні, стиснуті дані зберігаються у відокремлених блоках згідно формату словникового стиснення DEFLATE [1, 3]. Відповідно до цього формату, у стиснутих блоках містяться результати застосування до вхідного потоку алгоритму LZH [3], згідно з яким результати контекстно-залежного словникового алгоритму LZ77 [4] стискаються контекстно-незалежними кодами Хафмана.

Описуючи словникові алгоритми, фіксовану кількість попередніх закодованих неподільних елементів (літералів) вхідного потоку називають словником, а наступних незакодованих – буфером. Алгоритм LZ77 базується на заміні у вихідному потоці послідовності чергових літералів буфера посиланням на аналогічну послідовність літералів словника у вигляді пари чисел *<довжина; зміщення від закінчення словника>*. У випадку відсутності аналогічної послідовності літералів у словнику, перший літерал буфера переноситься у вихідний потік без змін. Після цього закодовані літерали переносяться з початку буфера в кінець словника і кодування продовжується аналогічно аж до закінчення літералів вхідного потоку. Наприклад, потік кодів біт пікселів *36 38 35 35 36 38 35 36 36 38 35 38 36 38 35 35 28* в закодованому вигляді записується як *36 38 35 35 <3; 4> 36 <3; 4> 38 <4; 12> 28*. Алгоритм LZ77 використовується у форматі PNG насамперед для зменшення надлишковостей між однаковими фрагментами зображення.

Під час декодування кодів алгоритму LZ77 окремі літерали копіюються у вихідний потік без змін. Пари ж *<довжина; зміщення>* декодуються шляхом послідовного копіювання з кінця вихідного потоку за вказаним зміщенням в кінець вихідного потоку необхідної кількості літералів. Природно, що алгоритм декодування має розрізняти окремі літерали та групи *<довжина; зміщення>*. Згідно з алгоритмом LZH, у форматі DEFLATE з цією метою довжини замін та окремі літерали кодуються разом числами в межах [0; 285]. При цьому числа з діапазону [0; 255] відповідають кодам окремих літералів, 256 позначає закінчення блоку, а числа з діапазону [257; 285] вказують на базові значення довжин. Після базових значень довжин міститься визначена форматом кількість біт, що разом з базовим значенням однозначно визначає довжину заміни. Зміщення зберігається після відповідної довжини аналогічно – у вигляді базового значення та додаткових біт. Базове значення зміщення знаходиться в межах [0; 29]. У форматі DEFLATE максимальне значення довжини закодованої послідовності може сягати 258, а зміщення – 32768.

Ідея використання контекстно-незалежних кодів Хафмана, що застосовуються для кодування окремих літералів і базових значень довжин та базових значень зміщень після виконання алгоритму LZ77, полягає у заміні чисел з більшою частотою (тут і надалі – абсолютною) кодами меншої кількості біт, ніж для чисел з меншою частотою. Коди Хафмана для літералів/довжин та зміщень визначаються для кожного блоку стиснутих даних, як правило, окремо (і називаються у цьому випадку динамічними), що сприяє покращенню стиснення. Автономне кодування Хафмана в форматі PNG насамперед зменшує надлишковості між переважаючими яскравостями пікселів зображення.

Згідно теореми Шеннона, елемент s_i з ймовірністю появи $p(s_i)$ найвигідніше кодувати $-\log p(s_i)$ бітами (тут і надалі логарифм береться за основою 2). Тоді середня довжина коду елемента після застосування контекстно-незалежного алгоритму має наближатися до *ентропії джерела* [3]:

$$H = -\sum_i p(s_i) \times \log\left(p(s_i)\right). \qquad (1)$$

Ентропія джерела зменшується при збільшенні нерівномірності розподілу ймовірностей між елементами. Середня довжина коду Хафмана співпадає з ентропією джерела лише тоді, коли для всіх елементів s_i довжини їх оптимальних кодів $-\log p(s_i)$ цілі. Чим більше $-\log p(s_i)$ відхиляються від цілих чисел, тим більшою стає різниця між середньою довжиною коду Хафмана і ентропією джерела, але ця різниця не перевищує одного біта.

Зменшити ентропію джерела при обробці зображень у форматі PNG намагаються за допомогою предикторів. *Предиктор* – це функція, що прагне, використовуючи значення відомих суміжних елементів, спрогнозувати (змоделювати) значення чергового елемента. Якщо піксел зображення характеризується декількома компонентами (наприклад, R, G, B), то предиктор кожної компоненти прогнозує значення згідно відповідних компонент сусідніх пікселів. В процесі використання цієї технології обчислюють і надалі кодують відхилення чергової компоненти від прогнозованого предиктором значення. Тому, у загальному випадку, процес застосування предикторів до кожної компоненти піксела у вузлі *(i, j)* можна записати формулою

$$\Delta_{ij} = C_{ij} - predict_{ij}, \qquad (2)$$

де C_{ij} – значення компоненти до застосування

предиктора, Δ_{ij} – значення компоненти після застосування предиктора, $predict_{ij}$ – значення предиктора, обчисленого для обраної компоненти.

Оскільки сусідні пікселі зображення мають, як правило, близькі кольори і тому близькі значення відповідних компонент, то часто значення предиктора буде співпадати зі значенням чергового елемента, найчастіше – буде близьким до цього значення і рідко – значно відрізнятиметься від нього. Тобто більшість значень Δ_{ij} будуть близькими до 0. Такий перерозподіл частот значень (а, отже, і ймовірностей) значно підвищує нерівномірність розподілу [5] і тому зменшує ентропію джерела (згідно (1)), а, отже, і довжину закодованої послідовності. Таким чином, предиктори використовуються, насамперед, для зменшення надлишковостей між сусідніми пікселами зображення, що мають близькі кольори. Крім цього, застосування алгоритму LZ77 та кодування Хафмана до результатів дії предикторів дозволяє зменшити надлишковості між однаковими змінами кольорів.

У стиснутих блоках формату PNG даним кожного рядка передує окремий байт, що визначає предиктор, який застосовується до компонент всіх його пікселів. На сьогодні форматом передбачено п'ять можливих значень цього байта [1], що визначає чотири різні предиктори: 0 – дані рядка не обробляються предикторами; 1 – предиктор рівний значенню відповідної компоненти зліва; 2 – предиктор рівний значенню компоненти зверху; 3 – предиктор рівний середньому арифметичному значень зліва та зверху; 4 – предиктор Піфа, що прогнозує значення у напрямку найменшого приросту. Опис предикторів формату PNG, міститься в [1], класифікація цих та інших предикторів наведена в [5].

На сьогоднішній день у спеціалізованому програмному забезпеченні для покращення стиснення зображень у форматі PNG найчастіше використовують метод перебору різних варіантів предикторів, розмірів блоків стиснутих даних та стратегій стиснення. **Метою** ж цієї **статті** є опис алгоритму мінімізації розміру стиснутих блоків динамічних кодів Хафмана за допомогою аналізу ефективності замін алгоритму LZ77.

2. АЛГОРИТМ ГЕНЕРУВАННЯ АЛЬТЕРНАТИВНИХ СТИСНУТИХ БЛОКІВ, ВИБОРУ НАЙКОРОТШОГО БЛОКУ З АЛЬТЕРНАТИВНИХ ТА ІТЕРАТИВНОГО ЗМЕНШЕННЯ ЙОГО РОЗМІРУ

Очевидно, що заміни LZ77 слід вважати ефективними, якщо вони записуються не більшою кількістю біт, ніж окремі літерали, які вони замінюють. Оскільки окремі літерали і базові значення довжин та зміщення у форматі DEFLATE записуються кодами Хафмана, то заміна j довжини len_j, що виконується починаючи з літерала s_k за зміщення $offset_j$, ефективна лише тоді, коли

$$\sum_{i=0}^{len_j-1} l_{s_{k+i}} \geq l_{len_j} + d_{len_j} + \lambda_{offset_j} + \delta_{offset_j}, \quad (3)$$

де l_m – довжина коду Хафмана літерала/довжини заміни m, d_m – кількість додаткових біт для запису довжини заміни m, λ_m – довжина коду Хафмана зміщення m, δ_m – кількість додаткових біт для запису зміщення m. Але короткі заміни виявляються, як правило, ефективними для стиснутих блоків, в яких інші заміни такої ж довжини теж враховуються. Це пов'язано з тим, що врахування замін однакової довжини може суттєво підвищити частоту цієї довжини заміни в розподілі літералів/довжин і, як наслідок, зменшити довжину її коду Хафмана. З іншого боку, врахування сукупності коротких замін може суттєво зменшити частоти окремих літералів і тому збільшити довжини їх кодів Хафмана. Крім цього, на довжини (загальні кількості біт) стиснутих блоків суттєво впливають зміщення між однаковими фрагментами вхідного блоку даних, адже більші зміщення кодуються більшою кількістю додаткових біт.

У форматі DEFLATE мінімальна довжина кодів Хафмана рівна 1, максимальна – 15, максимальна кількість додаткових біт довжини складає 5, а зміщення – 13. Тому, згідно (3), для будь-яких розподілів частот ефективними будуть заміни довжиною від 48 літералів. На практиці зображення, в яких окреме значення яскравості повторюється частіше від інших значень яскравостей разом узятих, зустрічаються надзвичайно рідко (хіба що однотонні заливки), як наслідок – довжини кодів літералів майже завжди перевищують 1, тому ефективними будемо вважати заміни від 24 елементів. Ефективність коротших замін залежить не лише від літералів, які вони замінюють, а й загалом від

розподілу стиснутого блоку. Для чергового блоку даних характер розподілу його літералів/довжин і, тим більше, зміщень наперед визначити неможливо, тому **для кожного вхідного блоку даних ми пропонуємо такий алгоритм мінімізації розміру відповідного стиснутого блоку:**

1) **створити альтернативні стиснуті блоки з різними максимальними довжинами неврахованих замін;**
2) **обрати серед них найкоротший блок;**
3) **ітеративно зменшити його довжину;**
4) **використати сформований блок для зберігання даних у форматі DEFLATE PNG-файла.**

Програми, що використовуються на сьогодні для стиснення зображень у форматі PNG, або ж враховують всі заміни, або ж відкидають найкоротші заміни для всіх блоків. Ми ж, по суті, пропонуємо визначати мінімальний розмір врахованих замін для кожного блоку даних окремо так, щоб мінімізувати відповідні стиснуті блоки. Розглянемо тепер практичні аспекти реалізації лише перших трьох кроків цього алгоритму, оскільки четвертий крок полягає у звичайному записі сформованих кодів у вихідний файл:

1. Очевидно, що генерувати альтернативні стиснуті блоки доцільно лише для тих максимальних довжин неврахованих замін, які зустрічаються в розкладі LZ77 чергового блоку даних і можуть виявитися неефективними, тобто не перевищують 24. Враховуючи те, що довжина кожного стиснутого блоку складається з суми довжин закодованих літералів/довжин замін і зміщень та додаткових біт, загальну довжину альтернативного блоку з максимальною довжиною неврахованих замін j для кожного блоку даних будемо розраховувати за формулою:

$$L_j = \sum_{i=0}^{255} n_{ij} l_{ij} + l_{256,j} + \sum_{i=257}^{285} n_{ij}\left(l_{ij} + d_i\right) + \\ + \sum_{i=0}^{29} \eta_{ij}\left(\lambda_{ij} + \delta_i\right), \quad (4)$$

де n_i – частота елемента чи базової довжини i, η_i – частота базового зміщення i. Як видно з цієї формули, для обчислення довжини альтернативного стиснутого блоку не потрібно зберігати сам блок. Достатньо знати лише частоти елементів розподілів альтернативного стиснутого блоку та згенерувати коди Хафмана згідно цих розподілів. На практиці достатньо зберігати частоти розподілів з усіма врахованими замінами та перелік замін, що можуть виявитися неефективними, а для аналізу довжин інших альтернативних стиснутих блоків відкидати з цих розподілів частоти замін до визначеної довжини та враховувати відповідні літерали.

2. Визначати максимальну довжину неврахованих замін (а, отже, і індекс) найкоротшого стиснутого блоку з альтернативних будемо за формулою

$$indexMinBlock = \min\left\{k \mid L_k = \min_{j=2,\,24} L_j\right\}. \quad (5)$$

Прискорити визначення найкоротшого з альтернативних стиснутих блоків можна шляхом перебору максимальних довжин відкинутих замін не до 24, а лише до 9, адже довгі заміни, по-перше, зустрічаються доволі рідко, і, по-друге, неефективні довгі заміни можуть бути ліквідовані на третьому кроці алгоритму. Обчислювати довжину кожного альтернативного стиснутого блоку згідно (4) можна безпосередньо, після генерації відповідних кодів Хафмана. Прискорити ж виконання цих розрахунків можна за допомогою окремого зберігання додаткових біт замін (оскільки вони не залежать від розподілів частот) та обчислення розміру блоків кодів Хафмана літералів/довжин та зміщень лише за абсолютними частотами їх елементів, як це описано в наступному розділі. Наведемо фрагмент програми мовою С, що реалізує два розглянуті кроки алгоритму:

```
// розносимо частоти замін та відповідних їм
// літералів, що можуть виявитися неоптимальними
for (i=0; i<countAnalizZamina; i++)
 {len = lenZamina[i]; offset=offsetZamina[i];
  // визначаємо базові значення та додаткові біти
  LengthToCode (len, code, extra, value);
  DistanceToCode (offset, codeD, extraD, valueD) ;
  poz=pozImageZamina[i]; // початок заміни в даних
  nayavnoLenZamina[len]=true; // розподіл наявний
 // фіксуємо параметри заміни для аналізу відкидань:
 for (j=0; j<len; j++) // реєструємо літерали,
  freqProbaLength[len][imageData[poz+j]]++;
 freqProbaLength[len][code]++; // базу довжини,
 freqProbaDistance[len][codeD]++; // базу зміщення,
 plusBit[len]+=extra+extraD;} // додаткові біти заміни
// встановлюємо ознаку наявності розподілу з усіма
nayavnoLenZamina[2]=true; // можливими замінами
// встановлюємо індекс попереднього наявного
pprNayavno=0; // альтернативного стиснутого блоку
// цикл по максимальних довжинах відкинутих замін
for (k=2; k<minLenZaminaLiteral; k++)
 if (nayavnoLenZamina[k]) // є заміни чергової довжини
  { // накопичуємо частоти літералів відкинутих замін
   for (i=0; i<=256; i++)
    {freqProbaLength[k][i]+=
      freqProbaLength[pprNayavno][i];
    // сумуємо частоти літералів розподілу з усіма
    // замінами та  частоти літералів відкинутих замін
    masAnalizLL[i]=freqCodeLengthTable[i]+
      freqProbaLength[k][i]; }
```

```
// накопичуємо частоти базових значень довжин
for (i=257; i<=285; i++) // відкинутих замін
 {freqProbaLength[k][i]+=
   freqProbaLength[pprNayavno][i];
  // обчислюємо різниці частот баз довжин розподілу
  // з усіма замінами та частот баз відкинутих замін
  masAnalizLL[i]=freqCodeLengthTable[i]-
   freqProbaLength[k][i]; }
// підраховуємо довжину розподілу літералів/замін
// для чергового альтернативного стиснутого блоку
sizeCode=countBitHufPackFreq(masAnalizLL, 286);
// накопичуємо частоти базових значень зміщень
for (i=0; i<=29; i++) // відкинутих замін
 {freqProbaDistance[k][i]+=
   freqProbaDistance[pprNayavno][i];
  // обчислюємо різниці частот баз зміщень
  masAnalizD[i]=freqDistanceLengthTable[i]-
   freqProbaDistance[k][i]; }
// додаємо довжину розподілу замін чергового
// альтернативного стиснутого блоку
sizeCode+=countBitHufPackFreq(masAnalizD, 30);
// накопичуємо додаткові біти відкинутих замін
plusBit[k]+=plusBit[pprNayavno];
// відкидаємо додаткові біти з обчисленої довжини
sizeCode-=plusBit[k];
if (sizeCode<minSizeCode) //черговий блок коротший
 {minSizeCode=sizeCode;//запам'ятовуємо цей
розмір
  // запам'ятовуємо максимальну довжину
  // відкинутих замін найкоротшого з розглянутих
  indexMinBlock=k; } // альтернативних блоків
// запам'ятовуємо індекс попереднього наявного
pprNayavno=k; } //альтернативного блоку
```

3. Після визначення максимальної довжини неврахованих замін найкоротшого з альтернативних стиснутих блоків для подальшого зменшення його розміру необхідно спочатку розрахувати довжини кодів Хафмана окремих елементів розподілів літералів/довжин та зміщень. З цією метою потрібно:

- згенерувати частоти розподілів найкоротшого з альтернативних стиснутих блоків, заміщуючи для замін, що не враховуються, в частотах розподілів з усіма замінами базові значення довжин та зміщень відповідними літералами;
- на основі сформованих частот розподілів розрахувати довжини кодів Хафмана літералів та базових значень довжин і зміщень;
- присвоїти елементам розподілів, що не використовуються, довжину коду Хафмана одиничного елемента, тобто максимальну довжину, оскільки такі елементи можуть з'явитися під час подальшого зменшення розміру найкоротшого з альтернативних блоків.

Зменшити довжину обраного стиснутого блоку можна за рахунок відкидання неефективних врахованих та врахування ефективних відкинутих замін згідно (3). Але врахування ефективних замін зменшує частоти окремих елементів і, відповідно, може збільшити після перерахунку довжини їх кодів Хафмана. Тому серед замін, неефективних з попередніми кодами Хафмана, можуть виявитися ефективні з перерахованими такими кодами. І навпаки: відкидання неефективних замін збільшує частоти окремих елементів та, відповідно, може зменшити після перерахунку довжини їх кодів Хафмана, тому серед замін, ефективних з попередніми кодами Хафмана, можуть виявитися неефективні з перерахованими такими кодами. Ось чому для обраного найкоротшого блоку з альтернативних у випадку, коли він не враховує всі заміни, доцільно після аналізу ефективності замін, коротших 24, перерахувати довжини кодів Хафмана елементів розподілів та ітеративно проаналізувати ефективність цих замін ще раз. Додаткові ітерації для перевірки ефективності замін недоцільні, оскільки вони суттєво не впливають на довжину обраного стиснутого блоку та значно сповільнюють виконання алгоритму.

3. ОБЧИСЛЕННЯ РОЗМІРУ БЛОКУ КОДІВ ХАФМАНА ЗА АБСОЛЮТНИМИ ЧАСТОТАМИ ЙОГО ЕЛЕМЕНТІВ

У форматі DEFLATE для кожного стиснутого блоку генерується два блоки кодів Хафмана – для літералів/довжин та для зміщень. Визначення розміру таких блоків дозволяє обрати найкоротший стиснутий блок з альтернативних.

Наближено обчислити розмір таких блоків можна за допомогою ентропії, до якої наближається середня довжина кодів Хафмана [3]. Справді, нехай кожен з елементів S_i зустрічається N_i разів в послідовності довжиною $N = \sum_i N_i$. Тоді $p(s_i) = N_i / N$ і загальна довжина закодованих даних, враховуючи (1), наближається до значення

$$N \times H = N \log(N) - \sum_i N_i \log(N_i). \quad (6)$$

Мовою С підпрограма для такого наближеного обчислення розміру блоку кодів Хафмана має вигляд:

```
// masFreq – масив частот елементів
// countAllFreq – загальна кількість частот в масиві
size=0; n=0;
for (i=0; i<countAllFreq; i++)
 {n+=freq[i]; // сумуємо частоти
  if (freq[i]>1) // для коректності обчислень
   size+=freq[i]*log(freq[i]); }
if (n) size=(n*log(n)-size)/log(2);
else size=0;
```

Для точного обчислення розміру блоку кодів Хафмана скористаємося ітеративним принципом їх генерації [3]: серед всіх частот шукаються дві найменші і надалі розглядається їх сума (породжена вершина); після поєднання всіх частот в одну і формування відповідного дерева поєднань в зворотному порядку утворюються коди елементів за допомогою дописування до коду вершини одному з поєднаних елементів одиниці, а іншому – нуля (коренева вершина не кодується, якщо дерево містить хоча б два елементи). Тобто **внаслідок поєднання двох найменших частот довжина блоку кодів Хафмана збільшується на суму цих частот**. Цей принцип дозволяє ітеративно обчислювати розмір таких блоків без генерації самих кодів елементів: до отримання однієї частоти серед всіх додатних частот знаходимо дві найменші, збільшуємо довжину блоку на суму цих частот і надалі замість цих двох частот розглядаємо їх суму. Звичайно, знаходити в масиві частот щоразу дві найменші недоцільно, адже це значно сповільнить обчислення. Тому відсортуємо додатні частоти за спаданням і будемо поєднувати дві останніх частоти. Крім цього, сума двох чергових частот не може бути меншою суми двох попередніх поєднаних частот, оскільки щоразу обираються два найменших значення. Отже, чергова сума поєднання може бути вставлена перед попередньою сумую частот, що значно скорочує область пошуку позиції її вставки у відсортований масив. Наведемо фрагмент програми мовою С, що реалізує обчислення довжини блоку кодів Хафмана з використанням описаних підходів і бінарного пошуку позиції вставки частот у відсортований масив (тут ділення на 2 замінено побітовим зсувом):

```
// сортуємо частоти масиву за спаданням
for (i=0; i<countAllFreq; i++)
 if (masFreq[i]>0) // елемент в розподілі наявний
  {element=masFreq[i];
   if (countFreq==0) j=0;
   else // бінарний пошук позиції для вставки
    {minIndex=-1; maxIndex=countFreq;
     while (maxIndex-minIndex>1)
      {j=(minIndex+maxIndex)>>1; // індекс середини
       if (masFreq[j]>=element) minIndex=j;
       else maxIndex=j; }
      j=maxIndex; }
   for (k=countFreq; k>j; k--) // зміщуємо менші частоти
    masFreq[k]=masFreq[k-1];
   masFreq[j]=element; // вставляємо знайдену частоту
   countFreq++; } // збільшуємо кількість частот, >0
if (countFreq==0) return 0; // якщо частоти відсутні
if (countFreq==1) // один елемент кодується бітом
 return masFreq[0];
countBit=0;
j=countFreq-2; // позиція для вставки поєднаних частот
while (countFreq>2)
 {// обчислюємо суму двох найменших частот
  element=masFreq[countFreq-1]+masFreq[countFreq-2];
  countBit+=element; // збільшуємо розмір блоку
  // шукаємо позицію для прямого включення суми
  while (j>0 && masFreq[j-1]<element) j--; // частот
  for (k=countFreq-2;k>j;k--)// зміщуємо менші частоти
   masFreq[k]=masFreq[k-1];
  masFreq[j]=element; // вставляємо суму частот
  countFreq--; } // зменшуємо к-ть необроблених частот
// кодуємо додатковим бітом дві найбільші частоти
return countBit+masFreq[0]+masFreq[1]; }
```

Прискорити розрахунок довжини блоку кодів Хафмана можна, насамперед, за рахунок врахування особливостей обробки окремих частот. Значну частину часу в процесі поєднання частот займає включення нової суми у відсортований масив, адже при цьому доводиться щоразу всі менші частоти від отриманої суми посувати вправо для формування позиції вставки. Пришвидшити виконання таких включень можна за допомогою використання однозв'язного списку, який реалізується додатковим масивом індексів більших елементів.

Значно ж прискорити розрахунок довжини таких блоків дозволяє врахування особливостей розподілу їх частот: як свідчать експерименти, біля 67 % частот малі (до 15) і часто повторюються. Вставка таких частот у відсортований масив і подальше опрацювання займає багато часу, хоча можлива їх окрема ефективніша обробка. Розглянемо, наприклад, розподіл, в якому частота 1 повторюється 9 разів. Відразу можна визначити, що в процесі обробки буде поєднано 4 пари частот зі значенням 1 і внаслідок цих поєднань розмір блоку кодів збільшиться на 8 та буде створено 4 частоти зі значенням 2. Дев'ята частота зі значенням 1 поєднається з більшою частотою. Тому **для частот до 30 включно підрахуємо кількості їх повторень (частоти частот) в окремому масиві, поєднаємо малі частоти до 15 включно з використанням цього ж масиву та запишемо у відсортований масив частот лише результати цих поєднань**. Така оптимізація дозволяє зменшити кількість елементів у відсортованому масиві частот на 30-40 %. Використання ж двох описаних модифікацій дозволяє в середньому точно розраховувати довжину блоку кодів Хафмана не повільніше наближеної оцінки розподілу з використанням ентропії. Фрагмент програми мовою С, що реалізує всі описані модифікації, має вигляд:

```
// сортуємо в масиві частоти, більші 30, за спаданням;
// для менших частот підраховуємо їх кількість
for (i=0; i<countAllFreq; i++)
 if (masFreq[i]>0) // елемент в розподілі наявний
```

```
  if (masFreq[i]<31)
   {//підрахунок кількостей частот до 30 включно
    masCountMinFreq[masFreq[i]]++;
    countMinFreq++; } // к-ть мінімальних частот
  else // вставляємо частоту у відсортований масив
   {element=masFreq[i];
    if (countFreq==0) j=0;
    else// бінарний пошук позиції для вставки
     {minIndex=-1; maxIndex=countFreq;
      while (maxIndex-minIndex>1)
       {j=(minIndex+maxIndex)>>1; // індекс середини
        if (masFreq[j]>=element) minIndex=j;
        else maxIndex=j; }
      j=maxIndex; }
    for (k=countFreq;k>j;k--) // зміщуємо менші частоти
     masFreq[k]=masFreq[k-1];
    masFreq[j]=element; // вставляємо знайдену частоту
   countFreq++; } // збільшуємо кількість частот, >30
if (countFreq+countMinFreq==0)
 return 0; // частоти відсутні
countBit=0;
if (countMinFreq>0) // частоти до 30 наявні
 {//аналізуємо частоти до 15 включно
  for (i=1; i<=15; i++)
nextElement:
   if (masCountMinFreq[i]>0) // наявна кількість частот
    {if (masCountMinFreq[i]>1) // є однакові частоти
      {// знаходимо кількість пар однакових частот
       countParaFreq=masCountMinFreq[i]>>1;
       // враховуємо поєднання пар в довжині блоку
       countBit+=(countParaFreq<<1)*i;
       // поєднуємо пари однакових частот
       masCountMinFreq[i<<1]+=countParaFreq; }
     // залишилася одна частота – шукаємо їй пару
     if (masCountMinFreq[i] & 1)
      {element=i++;
       while (masCountMinFreq[i]==0 && i<=15) i++;
       if (i>15) //пару не знайдено – завершуємо аналіз
        {i=element; masCountMinFreq[i]=1; break; }
       // пару знайдено – враховуємо в довжині блоку
       countBit+=element+i;
       masCountMinFreq[element+i]++;//сумуємо частоти
       masCountMinFreq[i]--; // одну частоту опрацювали
       // переходимо до більшої частоти
       goto nextElement; }}
  //вставляємо поєднанні частоти в кінець масиву
  for (j=30; j>=i; j--)
   if (masCountMinFreq[j]>0)
    for (k=0; k<masCountMinFreq[j]; k++)
     masFreq[countFreq++]=j; }
if (countFreq==1) // один елемент кодується бітом
 return masFreq[0];
bottom=countFreq-1; // індекс найменшої частоти
for (i=bottom; i>0; i--)
 prev[i]=i-1; // список індексів більших частот в масиві
prev[0]=countAllFreq; // ознака закінчення списку
index=prev[prev[bottom]]; // позиція для поєднань
//циклічне поєднання двох найменших частот
while (prev[bottom]!=countAllFreq) //до однієї частоти
 {masFreq[prev[bottom]]+=masFreq[bottom];
  bottom=prev[bottom]; // посуваємо вершину списку
  // враховуємо поєднання в довжині блоку
  countBit+=masFreq[bottom];
  // якщо сумарна частота більша наступного елемента
  if (prev[bottom]<countAllFreq &&
      masFreq[prev[bottom]]<masFreq[bottom])
   {i=bottom;
    bottom=prev[bottom]; // переміщуємо вершину
    // шукаємо позицію для вставки поєднання частот
    while (prev[index]<countAllFreq &&
           masFreq[prev[index]]<masFreq[i])
     index=prev[index];
    // вставляємо суму частот в список
    prev[i]=prev[index]; prev[index]=i; index=i; }
  else index=prev[bottom]; }
return countBit; }
```

4. РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТІВ

На завершення розглянемо результати застосування описаного алгоритму використання альтернативних стиснутих блоків для компресії восьми різнотипних 24-бітних зображень стандартного набору файлів ACT у форматі PNG (завантажити їх TIFF-версії можна, наприклад, з http://compression.ru/arctest/act/act-files.html).

Тестування проводилося за допомогою програми з CD до [1], у яку були внесені такі модифікації:

- забезпечена можливість виходу зі словника в буфер під час кодування повторів;
- реалізований вибір предиктора для рядка пікселів зображення на основі співставлення ефективності стиснення зображення без попередньої обробки з оцінкою нерівномірності розподілу після дії кожного предиктора;
- запроваджений аналіз альтернативних стиснутих блоків і кількостей додаткових біт при записі зміщень;
- розмір блоків даних збільшений до 64 Кб та відкинуті допоміжні текстові блоки.

Результати тестування наведено в табл. 1, 2, де показником компресії файлів обрано коефіцієнт стиснення, виражений в bpp, тобто у кількості біт, що в середньому витрачаються для кодування одного піксела зображення. Дані тестування програми без застосування розглянутого алгоритму наведено в третьому, а з застосуванням – у четвертому стовпці. Крім цього, для порівняння ефективності стиснення у другому стовпці таблиць вказано результати тестування програми Microsoft Photo Editor 2000, яка не застосовує предиктори, а в п'ятому та шостому – результати популярної серед Web-дизайнерів програми OptiPng (http://www.optipng.sourceforge.net), яка генерує короткі PNG-файли за результатами відповідно стандартного та максимального перебору предикторів, розмірів стиснутих блоків та стратегій стиснення.

Таблиця 1. Коефіцієнти стиснення (bpp) файлів зображень набору ACT у форматі PNG після застосування різних варіантів програм

Bmp-файл	Photo Editor 2000	Miano без алгор.	Miano з алгор.	OptiPng стан-дарт	OptiPng макс.
Clegg	9.06	5.69	5.00	5.37	5.32
Frymire	1.78	1.64	1.64	1.63	1.63
Lena	22.85	15.92	14.45	14.48	14.48
Monarch	18.36	13.16	12.61	12.51	12.49
Peppers	21.97	13.98	12.92	12.95	12.92
Sail	20.73	15.88	15.88	15.97	15.80
Serrano	1.84	1.72	1.72	1.70	1.70
Tulips	21.61	14.65	13.86	13.84	13.82
Середній	**14.77**	**10.33**	**9.76**	**9.81**	**9.77**
Сукупний	**10.89**	**7.70**	**7.29**	**7.35**	**7.32**

Таблиця 2. Час стиснення (c) файлів зображень набору ACT у форматі PNG різними варіантами програм на комп'ютері з частотою 300 МГц

Bmp-файл	Photo Editor 2000	Miano без алгор.	Miano з алгор.	OptiPng стан-дарт	OptiPng макс.
Clegg	3	14	15	72	996
Frymire	4	18	18	78	684
Lena	2	6	7	14	289
Monarch	3	12	13	33	733
Peppers	2	7	8	16	408
Sail	3	11	11	21	486
Serrano	2	8	8	23	336
Tulips	3	10	11	26	562
Разом	**22**	**86**	**91**	**283**	**4494**

Як свідчать результати тестування, застосування описаного алгоритму покращило коефіцієнт стиснення на 2 – 6% для 63% зображень, які, як правило, є неперервно-тоновими, та не вплинуло на компресію решти файлів, хоча й сповільнило виконання програми в середньому на 7%. Крім цього, реалізація розглянутого алгоритму дозволила наблизитися до коефіцієнта стиснення програми перебору, а для деяких зображень і перевершити його, витрачаючи для компресії в 1.9 – 66 разів менше часу.

5. ВИСНОВКИ

1. Для підвищення ефективності стиснення даних у форматах, що послідовно використовують алгоритми декількох методів, слід враховувати взаємний вплив цих алгоритмів.
2. Підвищити швидкість розрахунку розміру блоку кодів Хафмана за частотами елементів можна не лише з допомогою швидких алгоритмів сортування та використання списків, а й шляхом окремого попереднього опрацювання малих частот.
3. Розглянутий алгоритм дозволяє суттєво покращити коефіцієнт стиснення більшості зображень за рахунок вибору для кожного блоку даних найкоротшого з альтернативних стиснутих блоків динамічних кодів Хафмана та ітеративного зменшення його розміру.
4. Описаний алгоритм не вимагає модифікації декодера чи програм перегляду зображень і за рахунок зменшення розмірів файлів лише прискорює їх роботу. Саме тому він може бути ефективно використаний для збереження даних у стандартах, що використовують формат словникового стиснення DEFLATE.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Миано Дж. Форматы и алгоритмы сжатия изображений в действии: учеб. пособ. / Дж. Миано. – М. : Триумф, 2003. – С. 249-318. – (Практика программирования).

[2] Шпортько О. В. Використання альтернативних блоків стиснутих даних у форматі PNG / О. В. Шпортько // Комп'ютерні науки та інформаційні технології: Матеріали третьої Міжнародної конференції CSIT'2008. – Львів: Видавництво ПП "Вежа і Ко", 2008. – С. 149-153.

[3] Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео / Д. Ватолин, А. Ратушняк, М. Смирнов, В. Юкин. – М. : ДИАЛОГ-МИФИ, 2003. – С. 17-106.

[4] Ziv J., Lempel A. A universal algorithm for sequential data compression / J. Ziv, A. Lempel // IEEE Transactions on Information Theory. – May 1977. – Vol. 23(3). – P. 337-343.

[5] Бредихин Д. Ю. Сжатие графики без потерь качества [Електронний ресурс] / Д. Ю. Бредихин. – 2004. – http: //www.compression.ru/download/articles/i_less/bredikhin_2004_lossless_image_compression_doc.rar

***Олександр Шпортько**, аспірант кафедри інформатики та прикладної математики Рівненського державного гуманітарного університету. Закінчив Рівненський державний педагогічний інститут у 1997 році за спеціальністю "Прикладна математика".*

Наукові інтереси: стиснення даних без втрат, теорія інформації, теоретичні основи програмування, проектування та розробка баз даних.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

A CHOICE OF THE SHORTEST OF THE ALTERNATIVE COMPRESSED BLOCKS OF THE DYNAMIC HUFFMAN'S CODES IN THE FORMAT OF PNG

Alexander Shport'ko

Department of Informatics and Applied Mathematics
Rivne State Humanitarian University, Ukraine
33028, Rivne, 12, S. Bandery street
E-mail:chportko@ukr.net, chportko@yandex.ru

Resume: *in the article the algorithm of generation of the alternative compressed blocks of dynamic Huffman's codes for every block of data, choice of the shortest compressed block of alternative and iterative diminishing of its size for the improvement of the compression of images in the format of PNG is offered. The methods of the estimation of the size of block of Huffman's codes according to the absolute frequencies of its elements are examined in details. The fragment of programs outlaying the language of C, which allows exactly to determine the size of the compressed block by principle of generation of dynamic Huffman's codes and separate working of small frequencies of elements, using for this purpose on the average not more time, than for the close estimation of size with the use of entroping. As experiments show, realization of the offered algorithm, allows to improve the indexes of the compression of the majority of images in the format of PNG in 2 – 6 %.*

Keywords*: lossless compression of images, dynamic Huffman's codes, format of graphic files of PNG.*

INTRODUCTION

The format of graphic files of PNG was created on October, 1, 1996 for the effective preserving of bitmapped images without losses after the company "Unisys" began to require paying for the using of format of GIF [1]. In this work the method of increase of indexes of compression of images in this format due to the generation and comparison of sizes of the alternative compressed blocks is described in details.

In PNG-files which are used nowadays, compressed data are saved in separate blocks according to the format of dictionary compression of Deflate [1, 3]. In accordance to this format, the compressed blocks contain results of he using of the input stream of predictors and algorithm of the LZH [3], according to which the results of context-dependent dictionary algorithm of the LZ77 [4] are compressed by the context-independent Huffman's codes.

Describing the dictionary algorithms, the fixed amount of the previous coded indivisible elements (literals) of the input stream is called a dictionary, but next uncoded – ones a buffer. The algorithm LZ77 is based on the replacement in the output stream of sequence of next literals of buffer by sending the similar sequence of literals of the dictionary as a pair of numbers *<length; displacement is from the end of the dictionary>*. In the case of absence of the similar sequence of literals in a dictionary, the first literal of buffer is transferred in the output stream began without changes. After this coded literals are transferred from the beginning of buffer to the end of the dictionary and coding continues in the same way up to the ending of literals of input stream. For the providing of synonymousness of decoding separate literals and base values of lengths are encoded together. With the purpose of increasing of indexes of compression for displacements and literals/lengths different optimum dynamic Huffman's codes are generated in every block of the compressed data.

1. ALGORITHM OF THE GENERATION OF THE ALTERNATIVE COMPRESSED BLOCKS, CHOICE OF THE SHORTEST BLOCK OF THE ALTERNATIVE AND ITERATIVE DIMINISHING OF ITS SIZE

Obviously, replacements of LZ77 are considered to be effective, if they are written down by not

greater amount of bats, than separate literals, which they replace. As separate literals and base values of lengths and displacements in the format of Deflate are written down by the codes of Huffman, the replacement of j length of len_j, which is implemented beginning with literal of s_k for displacement of $offset_j$ is effective only, when executed (3), where l_m is the length of Huffman's codes of literal/length of replacement of m, d_m is the quantity of the additional bats for the recording of length of replacement of m, λ_m – length of Huffman's code of displacement of m, δ_m is the quantity an amount of additional bats for the record of displacement of m. But short replacements appear, as a rule, effective for the compressed blocks in which other replacements of the same length are also taken into account in. It is connected with the fact that taking of replacements of identical length can substantially increase the frequency of this length of replacement in the distributing of literals/lengths and, as a result, decrease the length of its Huffman's codes. On the other side, taking into account of the aggregate of short replacements can substantially decrease frequencies of separate literals and as a result multiply lengths of their Huffman's codes. Besides, the displacement between the identical fragments of entry block of data substantially influence the length of the compressed blocks because greater displacements are encoded by greater quantity of additional bats.

According (3) to the standard of Deflate, for any distributing of frequencies there will be effective replacements not less than 48 literals. In practice, images in which the separate value of brightness repeats more frequent by than other values of brightness taken together, meet extraordinarily seldom (for example solid fillings), as a result – lengths of codes of literals almost always exceed 1, that is why effective are supposed replacements not less than 24 elements. The efficiency of shorter replacements depends not only on literals, which they replace, but also on the distribution of the compressed block in general. It is impossible to define the character of literals/length distribution in regular block and moreover, the distribution of displacement for every block. **That is why we offer the following algorithm of minimization of size in proper compressed block:**

1) **to create the alternative compressed blocks with different maximal lengths of not taken into account replacements;**
2) **to choose the shortest block among them;**
3) **decrease iteratively its length by casting aside of not effective and accounting of effective replacements according to (3);**
4) **to use the formed block for data storage in the format of Deflate PNG-file.**

2. RESULTS OF EXPERIMENTS

Let's examine the results of the application of the described algorithm of the use of the alternative compressed blocks for a compression, for example, of eight different type 24-bits images of standard set of ACT files of in the format of PNG. Testing was conducted with the help of the modified program from [1]. In fig. 1 image compressing factors are given, and in fig. 2 is the proper time of compression (s). Testing results of the program are given without application of the considered algorithm it is resulted in the third, and with application – in the fourth column. Besides, for the comparison of the efficiency of compression in the second column of tables the results of testing of "Microsoft Photo Editor 2000", which does not apply predictiors, are given and in the fifth and sixth there are results of the popular among Web-designers program OptiPng (http://www.optipng.sourceforge.net), which generates short PNG-files as a result of accordingly standard and maximal surplus of predictors, sizes of the compressed blocks and compression strategies.

As the results of testing show, application of the described algorithm increased an aspect ratio in 2 – 6% for 63% images, which, as a rule, are continuously toned, and did not influence the compression of the other files, although slowed implementation of the program on the average of 7%. Besides, the realization of the considered algorithm allowed to approximate to the aspect of the program of surplus ratio, for some images to surpass it, outlaying for a compression in 1.9 – 66 less time.

3. CONCLUSIONS

1. For the increase of the efficiency of data compression in formats which use the algorithms of a few methods consistently, we should take into account the mutual influence of these algorithms.
2. The considered algorithm substantially allows to improve compression factor of most images ratio due to choosing the shortest of the alternative compressed blocks of dynamic Huffman's codes and iterative diminishing of its size for every block.
3. The described algorithm does not require modification of decoder or programs of image viewing and due to decreasing the size of files only accelerates their work. For this reason, it can be effectively used for saving data in standards, which use the format of dictionary compression of Deflate.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

GPRS-BASED REMOTE SENSING AND CLIMATE CONTROL SYSTEM USING CMOD CPLD

Wael M. El-Medany

Department of Communications and Electrical Engineering,
Faculty of Engineering, Fayoum University, Egypt
Computer Engineering Department, Information Technology College,
University of Bahrain, 32038 Bahrain
Email: wmelmedany@itc.uob.bh, waelelmedany@gmail.com
Webpage: http://people.man.ac.uk/~mbgedwme/
Tel: +973-39764964

Abstract: *This paper presents the design and VLSI hardware implementation of a remote sensing system for humidity and temperature in real time. The remote monitoring of the system based on a web design by using GPRS (General Packet Radio Service) network. Since full custom ASIC design takes long time with high cost, programmable logic devices as a programmable ASIC is a better choice for rapid design process and reasonable prices. The design has been described using VHDL (VHSIC Hardware Description Language), and then implemented in hardware using CoolRunner2 CPLD from Xilinx to achieve low cost with rapid prototyping. The design has been simulated and synthesized using Xilinx ISE 6.2i software tools, then test in hardware level using Digilent Spartan 3 starter kit as a hardware tools. The design offers a complete, low cost, powerful and user friendly way of 24 hours real time monitoring system.*

Keywords: *VHDL, GPRS, VLSI, CPLD, Remote Sensing.*

1. INTRODUCTION

Programmable Logic Devices (PLDs) are extensively used in rapid prototyping and verification of a conceptual design and also used in electronic systems when the mask-production of a custom IC becomes prohibitively expensive due to the small quantity [1]. Many system designs that used to be built in custom silicon VLSI [2] are now implemented in Programmable Logic Devices. This is because of the high cost of building a mask production of a custom VLSI especially for small quantity [3]. With the rapid development of computer technology, the monitored control design of the central air-conditioning system is becoming the core of the design for building automation system [4] and [5]. Real-time monitoring provides reliable, timely information of petroleum product's status, important in taking decisions for petroleum production improvement. Evaluation of petroleum production systems is a time consuming and difficult process because it means performing visits to selected petroleum fields to be able to measure and register certain physical, chemical and biological characteristics of the petroleum production areas [6]. Microcontroller has been used for the design of remote sensing and climate control systems [7] and [8]. Field Programmable Gate Array (FPGA) is a better choice than microcontroller for cost effective design, since the use of FPGA will integrate most of the components in one single chip, which in turn reduce the size of PCB [9], [10], and [11]. The disadvantage with FPGA is that it is volatile, once the power is switched off, the desig will be erased. For remote monitoring, GSM has been used by sending SMS messages [12], [13], [14], and [15]. This research introduces a web-based 24-hours real time remote monitoring and climate control system with low cost for building automation and petroleum production systems. The advantages in the introduced system compared to others introduced in [4-15] is the use of GPRS for real time monitoring with low cost compared to the cost of sending SMS in case of GSM MODEM as well as using a 40-pines Coolrunner module "CMOD" with low cost, nonvolatile, and easy for hardware prototyping compared to the use of FPGA. The design has been synthesized and implemented using Xilinx ISE 6.2i.

2. SYSTEM ARCHITECTURE

The system design mainly consists of three main unites: the server, the PC and the remote unit, which in turn consists of four units; the controller that will be implemented in Xilinx CMOD coolrunner-2 CPLD, the sensor circuit, the cooling system, and the GPRS MODEM as shown in Fig. 1. The PC and server are located in the control center and the remote unit is located in the remote land, where humidity and temperature are measured. The three main units are communicated through the internet and the GPRS network. The main function of remote unit is continuously measure the temperature and humidity and compares the measured values with a threshold level, operates the cooling system and sends messages through GPRS network to the control center in case of high temperature or humidity exceeds the threshold level. The main subunit of the remote unit is the controller that has been designed using VHDL, and then implemented on Xilinx Coolrunner CMOD CPLD.

The controller has two components; the first component dealing with humidity measurements; and the second component dealing with the temperature measurements. Each component has its own inputs, the humidity input is coming from the humidity sensor, and the temperature input is coming from the temperature sensor. Both of the humidity and temperature sensors are connected to CPLD through an analog to digital converter (ADC) as shown in Fig. 2. The GPRS is connected to CPLD through RS232, using the standard serial communications port. The communications between GPRS and controller has been achieved by the design of a UART (Universal Asynchronous Receiver Transmitter). The UART has been designed using VHDL, and implemented on the CPLD. On the PC unit a web page has been designed using VB.net, for remotely accessing the system board and sensing the temperature or humidity, as well as sending commands to controller for adjusting the cooling system.

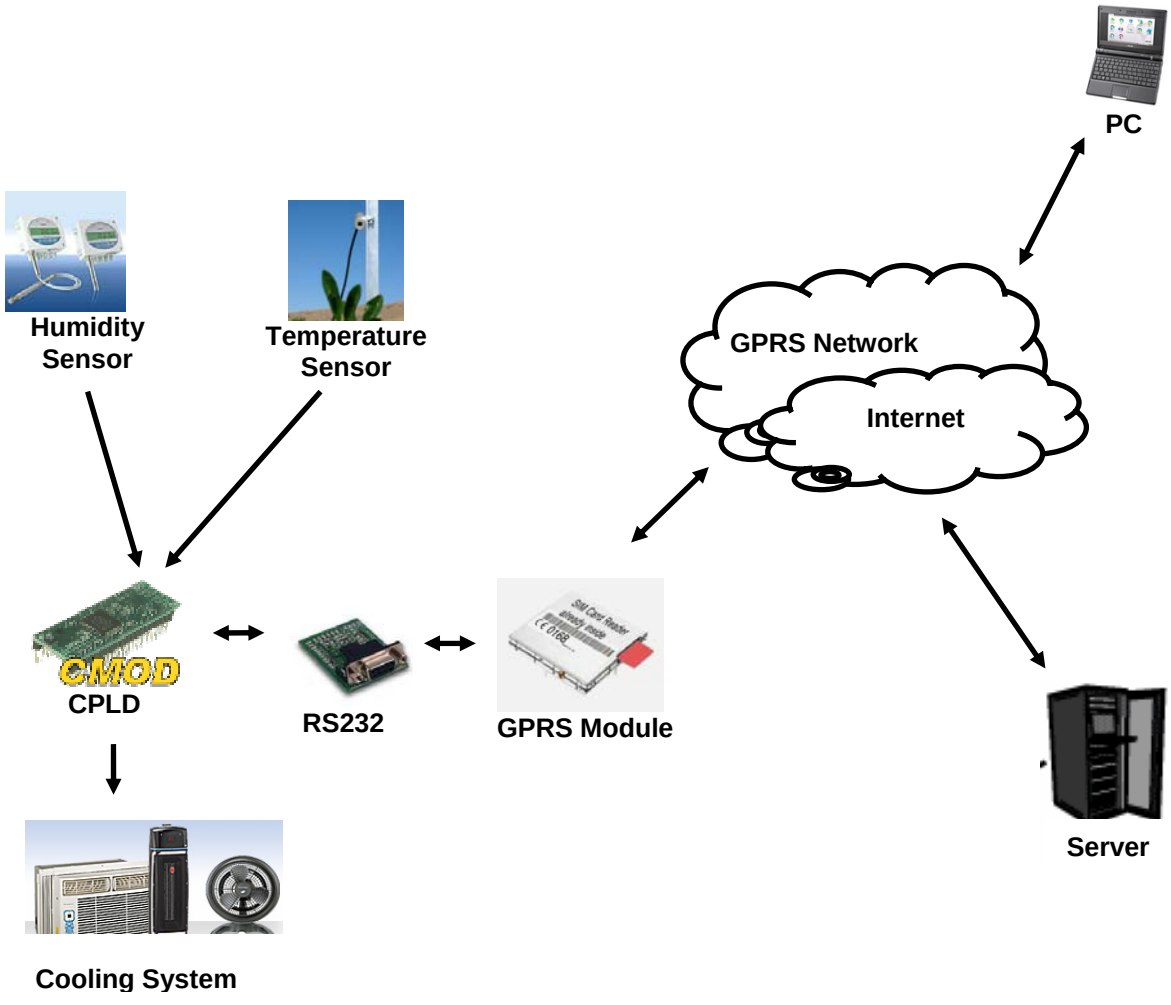


Fig. 1 – System Architecture

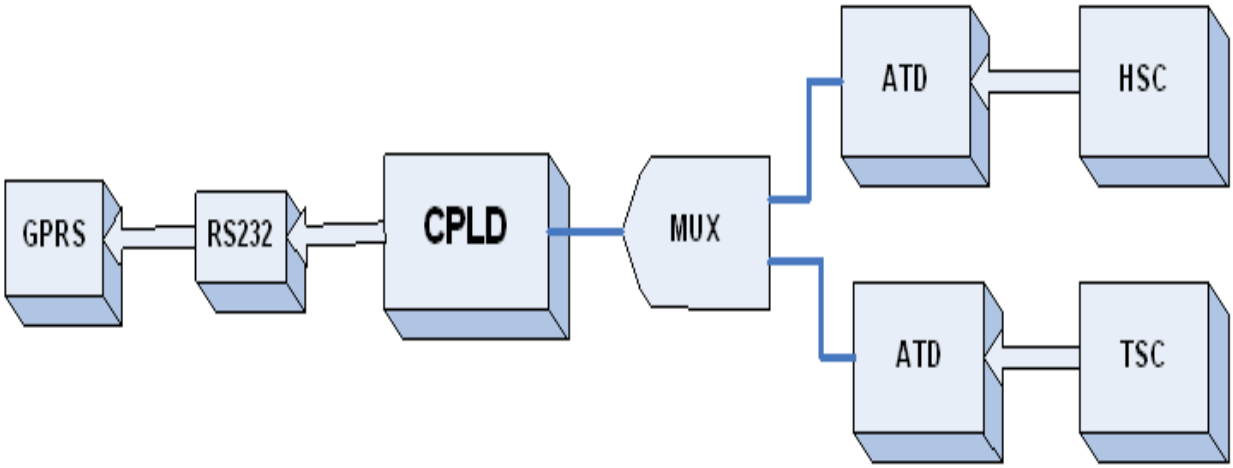


Fig. 2 – Block diagram for the main circuit board unit

3. STATE MACHINE CHARTS

This section concerned with hardware design of the controller circuit for the humidity and temperature sensing. Fig. 3 shows the State Machine (SM) Chart for the temperature and humidity control component that has been designed using VHDL and previously mentioned in section 2. For simplicity we have reduced the number of states to three states; initial state, HS (Humidity Sensing), and TS (Temperature Sensing) as shown in Fig. 3. In the initial state, controller checks for the variable "Sel" which select between temperature and humidity sensing. The HS state checks for high humidity, if it is true, the controller has to adjust the humidity and send message to the control centre, then the control goes back to the initial state. The same scenario happens in TS with temperature sensing. The value "Sel" that select between HS and TS, changes its value sequentially. By this way the controller is continuously monitoring the temperature and humidity status. By comparing the SM chart of fig. 3 with the block diagram of fig. 2, the input "Sel" in SM chart represents the "MUX" block in the block diagram. This multiplexer is used to save the number of used pines from CPLD by multiplexing both of the two reading from the humidity and temperature sensors.

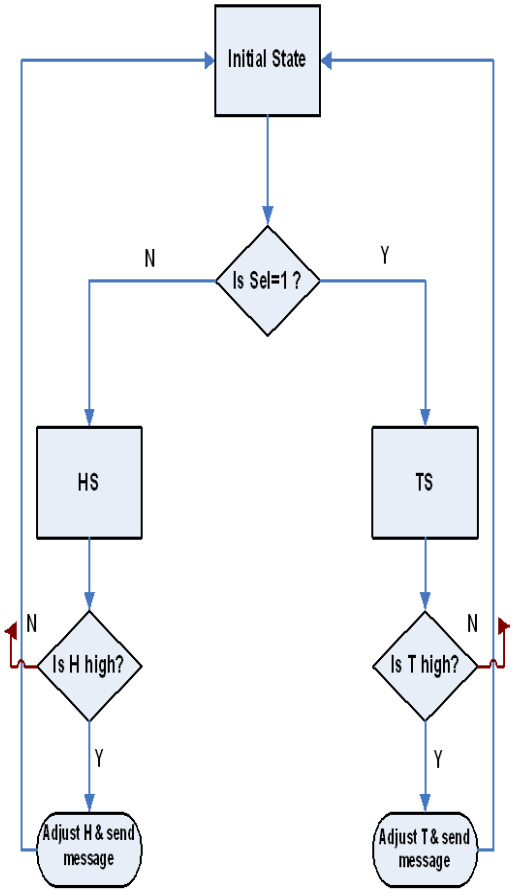


Fig. 3 – Simplified SM Chart for the sensing process

3. SIMULATION RESULTS

The VHDL models have been simulated functionally to verify the correctness of the behavioral description for the models. Fig. 4 shows the simulation results for the temperature sensing and control. Where “clk” is the system clock input, “TS” is an input bit vector of 8-bit which represents the inputs from the temperature sensor. The analog value of the temperature sensor has to be converted to digital value in binary representation using the ADC; this binary value is represented by “TS” input. The signals “SSCATH” is the cathode output for the seven segment display on the system board, and “AN” is the anode output for the seven segment, it is four bits that representing four digits, but only two of them are used. The values on “AN” represent the multiplexing between the seven segment digits, and the values on “SSCATH” represent the seven segment code for the decimal value. The signal “LSD” and “MSD” are internal signal that representing the two BCD (Binary Coded Decimal) digits of the temperature reading. Where “LSD” is least significant digit and “MSD” is most significant digit. Signal “T” is an output signal that indicates the high temperature and normal temperature. In fig. 5 the simulation results for the humidity sensing and control is given, where “PSEN” is reading value for humidity, “SHP” is the output signal that will high for high humidity, “S1” and “C” are the cathode and anode output for the seven segment respectively. Fig. 6 shows the simulation for the multiplexing between the humidity and temperature circuits. Where “TSC” and “HSC” are two 8-bit inputs from the analog to digital converter of the Temperature Sensing Circuit and Humidity Sensing Circuit, “Sel” is the selector input, and “Q” is an 8-bit output of the multiplexer. If (Sel=‘0’) then (Q = HSC), and if (Sel=‘1’) then (Q = TSC). The selector “Sel” is changing sequentially using binary counter. The MSB (Most Significant Bit) of TSC, HSC, and Q are on the left, whether the LSB (Least Significant Bit) are on the right.

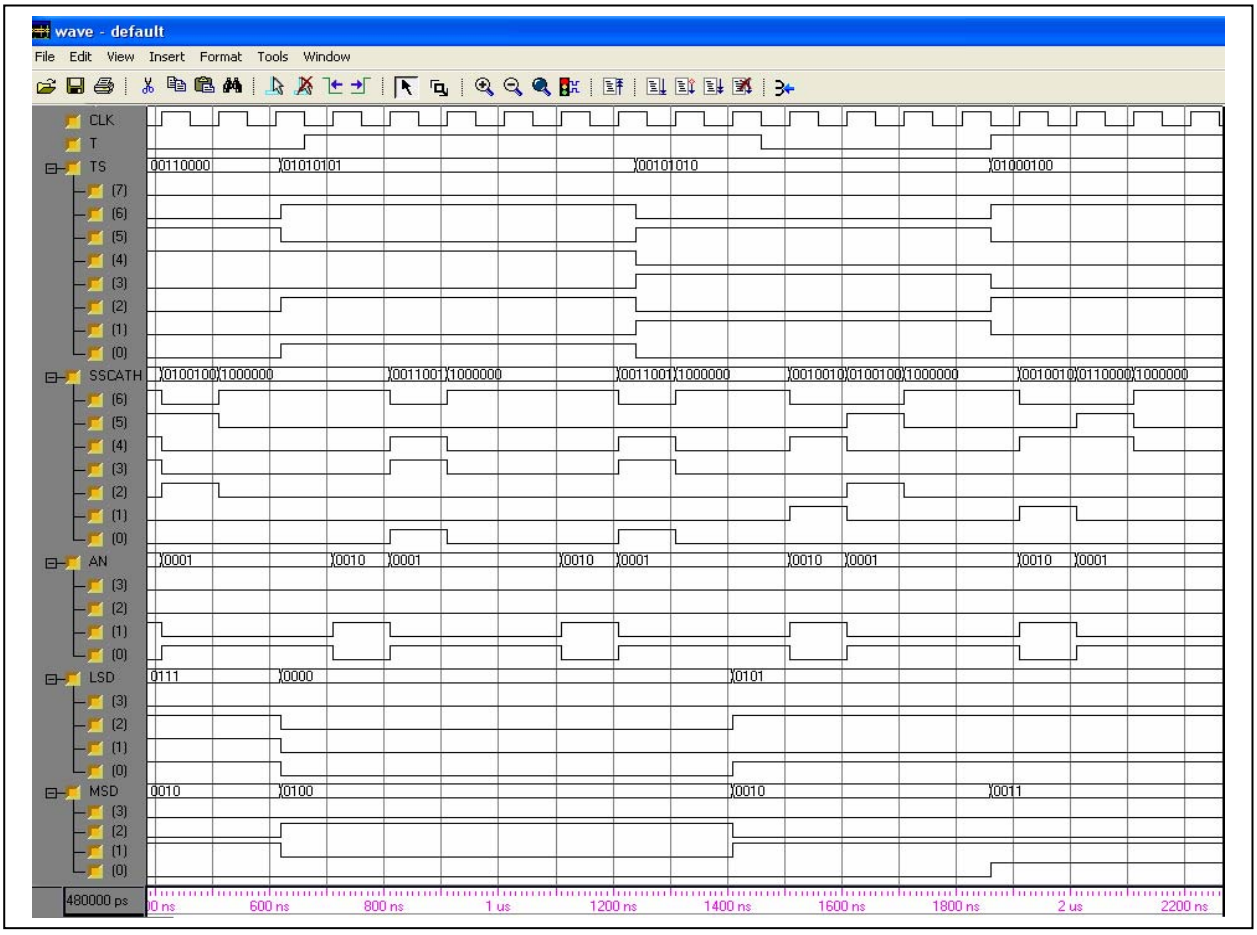


Fig. 4 – Simulation results for the temperature sensing and control

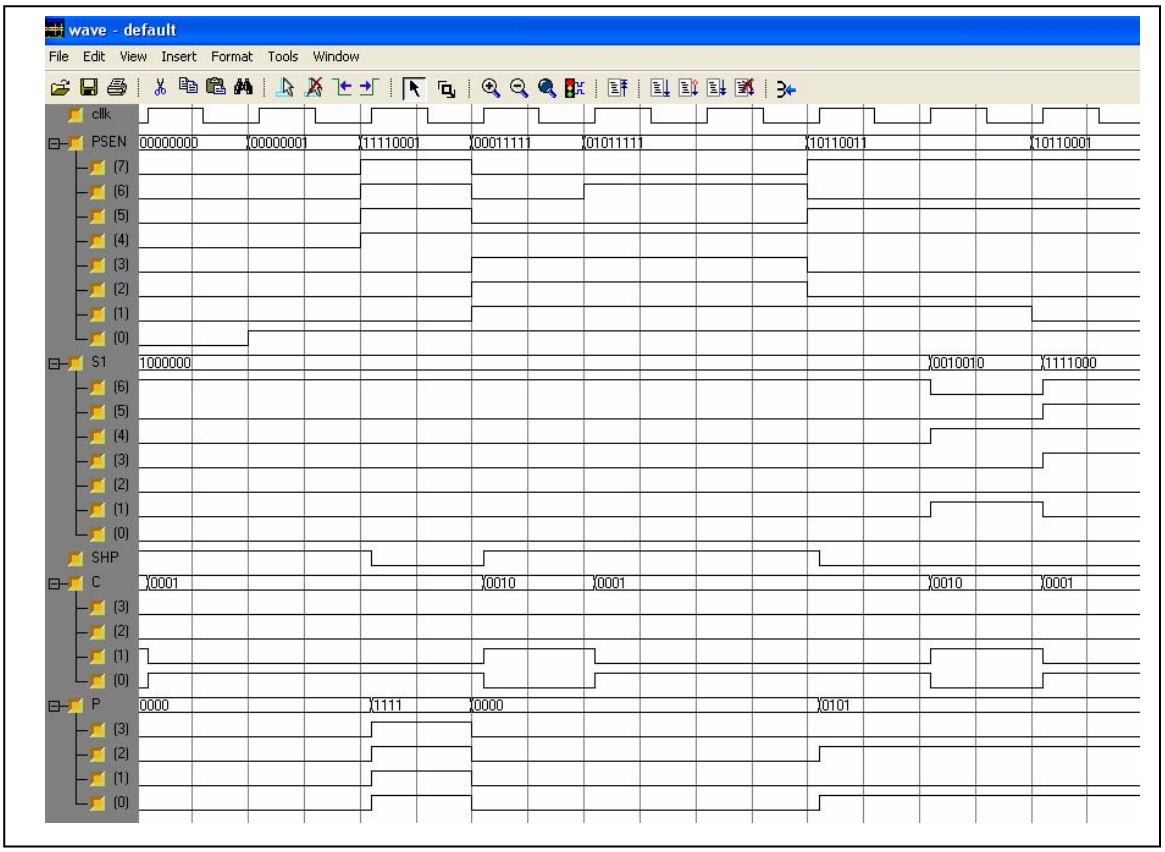


Fig. 4 – Simulation results for the humidity sensing and control unit

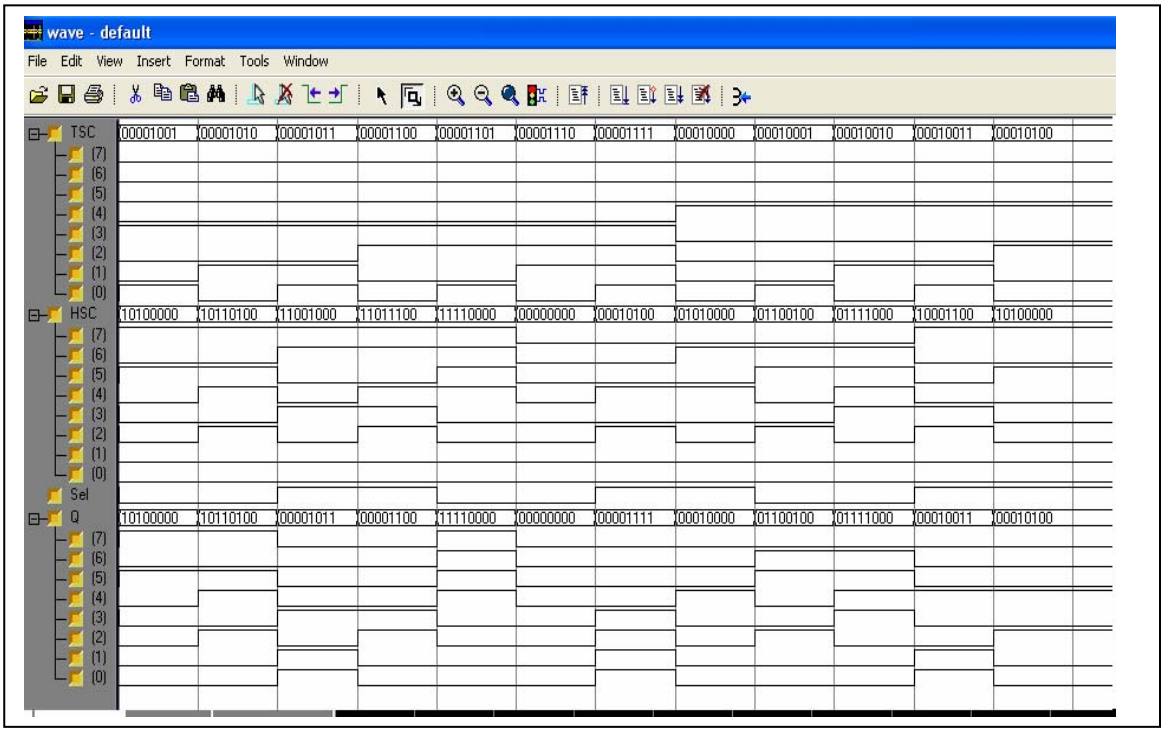


Fig. 5 – Simulation results for the multiplexing between temperature and humidity

5. CONCLUSION

A hardware design of a remote sensing system for humidity and temperature based on using GPRS for internet connection has been proposed in this article. The system was designed using VHDL in a high level design method. The system can remotely sense and control the temperature and humidity through the internet and GPRS network. All parts of the design have been tested in both simulation and hardware level. The implemented design targeted Xilinx Coolrunner CPLD CMOD with 40 pins for final prototype.

ACKNOWLEDGMENT

The author would like to thank Bader S Abdulla Taqi, Khalifa K Al-Temimi, and Khalifa A Rashed from Computer Engineering Department, IT College, University Of Bahrain for their valuable help.

6. REFERENCES

[1] Wayne Wolf, FPGA-Based System Design, Prentice Hall, 2005.

[2] Jan M. Rabaey, *Digital Integrated Circuits, A Design Perspective*, Second Ed., Prentice Hall, 2003.

[3] *Design of a VLSI Integrated Circuit*, IEEE, Piscataway, USA.

[4] Chen Xi, Chang Jinzhao, Liu Junfeng, "The monitored control design of the central air-conditioning system", In Control & Measurment 2007, vol.6, no.1, pp. 245-7.

[5] Jung-Ho Huh, M.J. Brandemuehl, "Optimization of air-conditioning system operating strategies for hot and humid climates" In Energy & Buildings 2008, vol.40, no.7, pp. 1202-13.

[6] Wu Ming-fang; Tang De-dong, "Smart instrument for measuring petroleum product's humidity based on electromagnetism oscillation technique", In Instrument Techniques and Sensor 2008, no.4, pp. 16-18.

[7] Ji Juan-zao, Shi Jun-yong, Zhang Ji-guang, "Applied controller of temperature and humidity by single-chip microcomputer", *In* Instrument Techniques and Sensor 2004, no.10, pp. 10-12.

[8] Li Xin, Qu Meng-ke, Rong Yuz, "Design of temperature and humidity fuzzy controller based on MSP430 MCU:, In Chinese Journal of Sensors and Actuators 2007, vol.20, no.4, pp. 805-8.

[9] W M El-Medany, "FPGA Implementation for Humidity and Temperature Remote Sensing System", 14th Annual IEEE International Mixed- Signals, Sensors, and Systems Test Workshop, IMS3TW 08, Canada.

[10] J. M. Jasso, G. O. Vargas, R. C. Miranda, E. V. Ramos, A. Z. Garrido, G. H. Ruiz, "FPGA-based real-time remote Monitoring system", Journal of Computers and Electronics in Agriculture, V 49, 2005, p 272–285.

[11] R C Miranda, E V Ramos, R P Vera, G H Ruiz, "Fuzzy Greenhouse Climate Control System based on a Field Programmable Gate Array", Biosystems Engineering, 2006, 94 (2), pp.165–177.

[12] GSM Based Remote Sensing and Control Systems Using FPGA, Wael M El-Medany, Mahmoud R El-Sabry, The IEEE International Conference on Computer & Communication Engineering, ICCCE'08, Kuala Lumpur, Malaysia, May 2008.

[13] G. Aranguren, L. Nozal, A. Blazquez, and J. Arias, "Remote control of sensors and actuators by GSM", IEEE 28th Annual Conference of the Industrial Electronics Society IECON 02, vol. 3, 2002 p 2306-2310.

[14] Wu, Bing-Fei, Peng, Hsin-Yuan; Chen, Chao-Jung "A practical home security system via mobile phones", WSEAS Transactions on Communications, v. 5, 2006, p. 1061-1066.

[15] Eddie M.C. Wong, "A Phone-Based Remote Controller for Home and Office Automation". IEEE Transactions on Consumer Electronics, Vo1.40, No.1, February 1994, pages 28-34.

Dr. Wael Mohamed El-Medany *is an assistant professor at the department of Communications and Electrical Engineering, Fayoum University, Egypt. Currently he teaches at the department of Computer Engineering, University Of Bahrain. He is IEEE and IEEE Communications Society member. He holds a PhD degree from Manchester University, UK, June 1999, with a major in Electrical Engineering and a minor in Chip Design and its application in error control coding. He got his MSc degree in computer communications from Faculty of Electronic Engineering, Menoufia University, October 1991, and his BSc degree in Electronic Eng from the same University May 1987. El-Medany also held teaching positions in Menoufia University, Cairo University, Obour Academy, Manchester University. His research interest in programmable ASIC design using VHDL.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

ПАРАДИГМА SEMANTIC WEB В КОНТЕКСТІ ЕЛЕКТРОННОЇ БІБЛІОТЕКИ: СЕРВІСИ ТА ІНТЕГРАЦІЯ ІНФОРМАЦІЇ

О.В. Новицький

Інститут програмних систем НАН України, проспект Академіка Глушкова 40, 03187, Київ, Україна.
alex@zu.edu.ua

Резюме: *В роботі представлено ряд теоретичних ідей та прикладні технології які можуть бути втілені при створенні семантичної електронної бібліотеки (ЕБ). Зокрема значна увага приділена, яким чином технологія Semantic Web використовується в різних аспектах ЕБ. Виділено основні рівні в структурі ЕБ які можуть бути носіями семантичного описання. Описані переваги такого підходу. Також висвітлено питання які постають при інтеграції класичних електронних бібліотек та місце Semantic Web в цих процесах. Зроблено короткий огляд провідних світових проектів з створення ЕБ з використанням технології Semantic Web.*

Ключові слова: *Semantic Web, електронна бібліотека, WSMO, семантична анотація.*

ВСТУП

Електронна бібліотека є складною інформаційною структурою. Саме поняття електронна бібліотека на даний момент конкретно не визначено. В роботах [1], [2], [3] було проведено аналіз стосовно цього поняття. Ми зупинимося на такому підході, що електронна бібліотека це об'єднання через мережу електронних текстів, документів, зображень, звуків, наукових даних та програмного забезпечення яке є ядром сьогоднішнього Інтернету, а в майбутньому через організацію доступу до електронних бібліотек буде утворюватися база знань людства. Цей підхід породжує так звану колективну пам'ять. Поняття колективної пам'яті саме для цифрових бібліотек виникло відносно недавно. Проте цей термін набув широкого поширення, зокрема комітет IEEE Technical Committee on Digital Libraries, трактує це поняття як сукупність електронних бібліотек, електронних музеїв, електронних архівів. Зазвичай основна інформація яка передавалася це була текстова інформація, про те на разі, передається інформації інших типів відео, звук, фотографії та ін.. [4].

Такий підхід є інноваційним оскільки дає цілісний доступ до інформації будь-яким користувачам будь-де. Розвиток сучасних інформаційних технологій дає можливість реалізувати це.

Оскільки сучасний науковий пошук пов'язаний з великою кількістю даних, тому важливо щоб цю інформацію могли використовувати всі науковці крім цього для наукових даних необхідним є прослідкування виникнення цих даних. Тим самим можна гарантувати їх достовірність, водночас зберігати унікальні екземпляри в бібліотеках, музеях та архівах.

Сьогодні в багатьох музеях та бібліотеках зберігається величезна кількість безцінної інформації, проблема в тому, що до неї є поки тільки фізичний доступ. Створення колективної пам'яті позитивно відобразиться на науці, такий підхід повинен стати серйозним поштовхом.

Відомо, що кожна наукова група збирає та поповнює свій інформаційний фонд, причому якщо кожна така група працює в одному напрямі то відповідно і інформаційні фонди будуть одного змісту, але можливо різної структури, що призводить до неефективної роботи, та різного тлумачення наукових понять.

Обмін інформацією дасть можливість аналізувати дані спостережень одночасно багатьом науковим групам навіть якщо вони будуть знаходитися на дуже великих відстанях. Таким чином утворюється спільний робочий простір, що дозволяє всім взаємовигідно працювати над проблемо.

Колективна пам'ять утворює так звані портали знань та контенту, що представляє собою мережу з розподіленими ресурсами.

Розвиток колективної пам'яті одночасно

потребує розвитку інших напрямів.

•Зберігання. Система Колективної пам'яті повинна та здатна зберігати великі об'єми інформації різнорідних форматів.

•Інтерфейс користувача. Один з найважливіших компонентів колективної пам'яті, який повинен представляти велику кількість сервісів, для взаємодії між користувачем та інформацією яку він шукає.

•Класифікація та індексація. Дає змогу групувати об'єкти. Однак виявлено, що на це сильно впливає індивідуальне сприйняття та великий обсяг інформації яку необхідно індексувати.

•Інформаційний пошук. В цій області існує багато методів пошуку, включаючи пошук мета даних та контенту. Визначити корисність результату пошуку може тільки сам користувач. Для покращення ефективності використовують додаткові метадані, які описують документ. Дослідники також зосереджуються на автоматизації створення і обслуговування параметрів користувача для використання їх в процесі пошуку.

•Адміністрування та збереження. Традиційні бібліотеки зберігають копію книги, музеї зберігають фізичний експонат. Система колективної пам'яті дозволяє зберігати декілька версій документа. Окрім того цифрова бібліотека може розмежовувати права доступу до авторських екземплярів тим самим зберігаючи авторське право. І всі перегляди будуть автоматично фіксуватися. Механізм захисту повинен бути надійним для виключення несанкціонованого доступу. Зміни технологій організації структури середовища зберігання інформації, доступу до неї та старіння засобів збереження становить серйозну проблему яка повинна також вирішуватися.

Окрім розглянутих вище напрямків необхідно також збільшувати степінь деталізації. Створення нових схем мета даних, зосередження на інформаційному вмісту а не на інформаційних об'єктах. [5].

Отже оперування такими обсягами інформації породжує певні проблеми. З погляду на вище сказане ці проблеми можна розділити на дві великі групи, перша група проблем пов'язана з технічними труднощами організації збереження інформації та доступу до неї, друга група проблем пов'язана з логічною організацію колективної пам'яті та забезпечення доступу до неї з подальшим аналізом змісту. Ця робота стосується вирішення проблем 2-гої групи.

1. ПАРАДИГМА SEMANTIC WEB В КОНТЕКСТІ ЕЛЕКТРОННОЇ БІБЛІОТЕКИ. ОГЛЯД СВІТОВИХ ПРОЕКТІВ

Використання семантичних технологій в електронних бібліотеках було приділено увагу в багатьох Європейських проектах:

В проекті SWHi [6] онтологія розроблена на основі електронної бібліотеки з точки зору, коли наші основні джерела даних в репозиторії описані метаданими. Це метадані відображається і зберігається в онтології, яка базується на онтології схеми. Крім того, буквені значення в метаданих, наприклад, заголовок, піддається аналізу в наслідок якого видобуваються імена сутностей, події та термінологія. Для збагачення онтології, також видобувають нову зв'язану інформацію з обраних веб-документів.

Пошук в цій системі реалізований у двох формах, простий та складний. Система використовує мову запитів RDF таку як SeRQL. Процесор генерування запитів SeRQL стикатися, принаймні з двома проблеми. По-перше, він не знає, в якому класі чи властивості можуть бути знайдені слова. Щоб уникнути цю проблему, прикладне програмне забезпечення Semantic Web, таке як OpenAcademia [13] вимагає від користувачів вводити ключові слова у відповідне поле (автор, назва або рік) в її розширений пошуковий інтерфейс. По-друге, існують деякі обмеження в підстроках відповідності SeRQL при використанні символу загальності '*'. Цю проблему можна вирішити за допомогою інформаційно-пошукових програм, таких як Lucene яка забезпечує потужний алгоритм, точного і ефективного пошуку.

Окрім самого пошуку, важливим також є питання представлення результатів пошуку. Одним із напрямків є візуалізація пошуку. У Semantic Web, візуалізація стає все більш важливою. Існують випадки складних взаємини між ресурсами, які не можуть бути представлені за допомогою простого списку. Крім того, як правило, відображається тільки невелику кількість результатів пошуку (в діапазоні 10-20 результатів на сторінці). До документів які знаходяться в хвості результату пошуку, швидше за все, ніколи не будуть зверталися.

Загальна архітектура системи SWHi показано на Рис. 1.

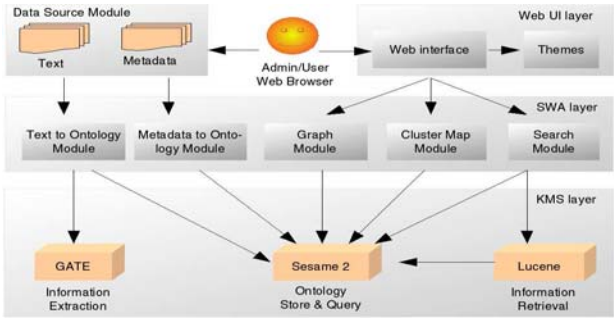


Рис. 1 – Архітектура SWHi

Для розвитку SWHi онтології, повторно використовуються наявні онтологічні ресурси для структурування і збереження історичної інформації, а саме: PROTON базова онтологія, таксономія предметної класифікації NewsBank/Readex, Дублінське Ядро та та словник FOAF Vocabulary. Це онтології зберігається з використанням Sesame2.

Проект eCulture є семантична пошукова система, яка дозволяє одночасно шукати в кількох колекціях установ культурної спадщини [7]. Це робиться шляхом перенесення цих колекцій в RDF шляхом зв'язування об'єктів колекцій як екземплярів класів через загальнодоступні словники, тим самим створюючи великий RDF граф.

Потім, в ході пошуку, цей граф трасується і деякі підграфи повертаються у вигляді результату.

Основним механізмом пошуку є використання Prolog.

Більшість Semantic Web додатків реалізовано за зразком реляційних баз даних, де прикладна логіка має доступу до бази даних на основі SQL [8]. Існує ряд еквівалентів SQL для Semantic Web, таких як SeRQL [9] і рекомендація W3C SPARQL [17]. Обидва дозволяють виразити граф, що складається з низки обов'язкових і необов'язкових ребер та вузлів, з розширеними умовами літеральних значень.

SeRQL відповідає вираженню графу на транзитивно закритому виразі використовуючи семантику RDFS. Стандарт SPARQL не визначає, чи виконується логічний наслідок, який виконаний механізмом судження СУБД.

Однак додатки eCulture, не використовують SeRQL або SPARQL. Замість цього, запити прикладної логіки виражаються як Prolog цілі на необроблених даних RDF та/або модулях судження RDFS/OWL.

Проект IPISAR (Image Preservation, Information Systems, Access and Research) досліджує розповсюдження, вивчення і раціональне використання культурної спадщини, та спроби представити вирішення загальних проблем в цих областях в рамках Semantic Web (SW) [11].

В рамках проекту розроблений додаток "Pescador", який буде зберігати каталогізовані дані в трійках які зберігатимуться в хранилищах (чиї функції будуть такі ж, що в реляційної бази даних в традиційних системах). Який як показала практика необхідно в подальшому вдосконалювати, зокрема забезпечення більш гнучких механізмів з подолання обмежень мови DSL яка була створена в рамках IPISAR.

Для досягнення цієї мети ми пропонуємо семантичну компоненту архітектуру (SCA), тобто, адаптація компонентів архітектури відповідно до принципів SW, в якому дані, структура та правила прикладної логіки, тісно пов'язані між собою.

SCA повинен координувати "компоненти", що підключаються, які б були б обгорнуті оболонкою яка б могла взаємодіяти з наступними типами: схемами; обмеженнями; правилами виводу; онтологіями; визначення шляхів; програмних кодом; специфікацією виводу; інформацією про конфігурацію Abox; посиланнями до зовнішніх джерел даних.

Алгоритм вилучення інформації з графів часто потребує визначення шляхів між ресурсами. Ці шляхи в значній мірі відрізняються по довжині і складності, тому SCA повинна включати засоби визначення моделі шляху. Попередній огляд існуючих механізмів визначення шляхів показує, що SPARQLeR розширення SPARQL може бути кращим кандидатом для адаптації для SCA і Пескадор. Автори [12] вважають, що підтримка семантичного шляху запиту повинна бути невід'ємною складовою RDF мови запитів. SPARQLeR (SPARQL extended with Regular paths), розширення мови запитів SPARQL, який додає підтримку для семантичного шляху запиту. Пропоноване розширення вписується в загальний синтаксис і семантику SPARQL і дозволяє легко формулювати запити за участю широкого кола регулярних шляхів в моделях RDF графів.

Проект EPOCH та AMA [13] представляють ЕБ як великий індекс покликаний служити в якості довідника для пошуку і повернення цифрової інформації яка зберігається на веб-сайті в різних форматах і архівах.

Перша проблема полягає в тому, що кожний довідник має свою пошукову систему і використовує свою граматику метаданих для опису та індексації даних, зокрема, що вона ніколи не буде працювати на інших системах. Жодна з цих систем метаданих може проаналізувати всю інформацію на веб-сайті,

якщо ми не будемо робити їх доступними через машину зрозумілій формі з використанням RDF [14].

Друга проблема стосується безпосередньо інформації: величезна різноманітних форматів, що використовуються для індексування даних, є великою перешкодою на шляху до інтеграції, і повинні бути серйозно проаналізовані. Навіть якщо ми обмежуємо наші зусилля виключно для культурної спадщини, архіви (наприклад, бази даних музеїв і колекцій, археологічні розкопки звіти, доповіді та інші неструктуровані дані), ми змушені визнати, що інформація, також є гетерогенною.

Щоб створити єдиний концептуальний шар, семантична інформація повинна бути взята з бази даних, HTML-сторінок, описових текстів, метаданих і повинна буди представлена в стандартному форматі, з метою отримання концептуального змісту інформації створивши концептуальний мапінг.

Як тільки концептуальний шар для даних і метаданих готовий, семантична інформація буде зберігатися в контейнері засновані на RDF і онтології.

Це реальним місцем інтеграції, де об'єднані основні відомості з різних електронних архів можуть бути переглянуті та в яких можливо здійснити пошук як в єдиній цілій електронній бібліотеці. RDF мова надійною і достатньо гнучкою, щоб забезпечити сумісність і забезпечити загальну основу не тільки для цифрових бібліотек, але і з інших систем та послуг.

Мапінг є одним із самих важливих кроків при інтеграції даних. Для спрощення та доступності процесу мапінгу в проекті AMA було розроблене програмне забезпечення AMA Mapping Tool гнучкий інструмент який сприяє мапінгу різних археологічних та музейних колекції моделей даних (з різною структурою, а також неструктуровані дані, тобто текстовий опис) на загальний стандарт ґрунтується на CIDOC CRM-онтології.

Аналізуючи ці проекти можна стверджувати, що існують ряд проблем при створенні електронних бібліотек та їх інтеграції з використанням семантичних технологій. Наприклад така проблематика як створення підходів до семантичної анотації електронних об'єктів. Особливо це стосується семантичної анотації контенту електронних бібліотек у випадку, якщо контент представлений у різних форматах та у різних галузях знань людства.

Одним із способів вирішення цієї проблеми може бути узагальнена формальна модель анотації.

Іншою проблемою яка постає при оперуванні великої кількості гетерогенної інформації це забезпечення відповідних сервісів. Оскільки сервіси є специфічними для різних форматів документів та повинні враховувати особливості вимог користувачів для обробки цієї інформації.

2. ФОРМАЛЬНА МОДЕЛЬ АНОТАЦІЇ

Для того щоб показати переваги семантичного підходу до електронної бібліотеки на відміну від класичної необхідно виділити ті структурні елементи які раніше не мали семантичної моделі, і до яких ми пропонуємо цю модель побудувати.

Основними двома компонентами електронної бібліотеки є її контент та набір програмного забезпечення для роботи з цим контентом. Для початку розглянемо контент ЕБ. Інформація в електронних бібліотеках описується в термінах електронні об'єкти (Digital objects – DO), які являють собою мультимедійний контент і метадані [15]. Оскільки обсяги DO значні, то для спрощення пошуку та класифікації використовують анотування DO.

Формальна модель, яка запропонована в [16], виділяє два підходи до розуміння анотацій: анотації як метадані або анотації як контент.

У першому випадку ми маємо справу з різноманітними схемами метаданих (Dublin Core, MARC і ін.) які використовуються для опису інформаційних ресурсів. Ці анотації насамперед направлені на користувача.

У другому випадку анотації як представлення контенту призначені для автоматизованої машинної обробки. Ці анотації надають семантику документа. Семантична анотація – анотація написана формальною мовою з добре визначеною семантикою і заснована на онтологіях. Фактично ці анотації є формальною моделлю DO, з можливістю машинної обробки.

Семантика контенту в свою чергу, може визначатися на основі зовнішніх зв'язаних онтологій, що дозволить будувати семантичну модель документу, де зв'язок визначається між окремими сегментами DO, та на основі семантики зв'язків між структурними компонентами DO, де зв'язок визначається між логічними закінченими структурними компонентами Рис. 2.

Модель, яка представлятиме цифровий об'єкт повинна відображувати фактичний зміст даного об'єкту. Серед безлічі розроблених моделей, ми використовуємо модель, яка запропонована в [15], [16] зі змінами та уточненнями, які враховують використання онтологій.

Розташування кожного цифрового об'єкту

також як і анотації ідентифікується унікальним ідентифікатором – посиланням (link). Окрім цього посилання сполучає цифровий об'єкт і анотацію, і може відображувати відношення між об'єктами. Отже, можна виділити два типи посилань:

посилання анотації (Annotate link) – відображує відношення в середині цифрового об'єкту, який може бути як документом, так і анотацією;

посилання відношення (Relate-to link) – визначає відношення зовнішнього цифрового об'єкту до об'єкту, що анотується.

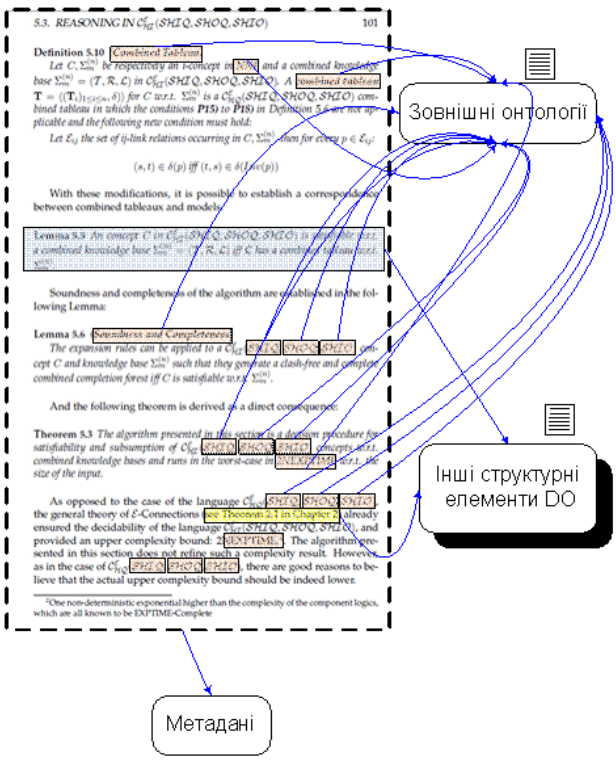


Рис. 2 – Анотування документа шляхом визначення відношень між термінами, лемами та зовнішніми отологіями а також визначення зв'язків між логічно закінченими елементами з іншими частинами DO. Окрім цього DO описується метаданими.

Нехай LT множина типів посилань, тоді LT містить наступні типи посилань $LT = \{AnnotateLink, Relate\text{-}toLink\}$.

Посилання надаються у вигляді ідентифікаторів. Загальноприйнятими ідентифікаторами можуть виступати URI, DOI, OPENURL, Persistent URL (PURL), PURL-based Object Identifier. $H(k)$ множина ідентифікаторів цифрових об'єктів у момент часу k.

У цій моделі цифровий об'єкт надається у вигляді потоку. Потік sm це кінцева послідовність:

$$DO(\Sigma) \to sm : I = \{1,2,...,n\},\ n \in \mathbb{N}$$

де e $= (j_1, j_2, ..., j_n)$ – алфавіт символів.

Якщо ми маємо потік sm :

$$DO(\Sigma) \to sm : I = \{1,2,...,n\},\ n \in \mathbb{N},$$

то в цьому потоці ми можемо виділити неперервний сегмент st_{sm} послідовності чисел a, b так що:

$$st_{sm} = [a,b],\ 1 \le a \le b \le n,\ n \in \mathbb{N}.$$

Безліч сегментів ми позначаємо ST , так що $\forall st_{sm_i} \in ST,\ i = 1,2,...,n.$

Якщо цифровий об'єкт DO має безліч унікальних ідентифікаторів H то функція hsm відображує унікальний ідентифікатор до сегментів які містяться в DO :

$$h \xrightarrow{hsm} st_{sm_i},\ n \in \mathbb{N}$$

Потік вимагає, щоб функція не мала властивостей сюрьїктивності і інєктивності. Кожен цифровий об'єкт може мати принаймні один потік.

SM визначає множину потоків, так що $\forall sm_i \in SM,\ i = 1,2,...,n$.

Анотацію можна розглядати як процес розширення онтології O_i. Розглянемо найпростіший випадок, коли розширення відбувається шляхом додавання нових екземплярів онтології O_i. Кожний клас онтології O_i будемо позначати через kl_i, а множину класів через KL.

Анотація $a \in A(k)$ це кортеж:

$$a = \begin{pmatrix} h_a \in H(k), \\ A_a \subseteq KL(k) \times LT \times ST(k) \times \\ \times SM(k-1) \times H(k-1) \end{pmatrix}$$

де h_a – унікальний власний ідентифікатор анотації a, тобто $h(h_a) = a$;

A_a множина n -арних відношень анотації a і визначається як добуток множин KL, LT, ST, SM та H.

У випадку анотування веб-документів, формальна модель зміниться. Нехай A множина всіх анотацій a, а D множина документів, відповідно, $DO = D \cup A$, причому підмножина множини DO позначатимемо do, тобто

$do \in DO$.

Анотацією веб-документа називатимемо мічений граф:

$$G := \left(\left(DO, E_{da} \subseteq A \times DO\right)\right),$$

де $DO = D \cup A$ вершини графа;

$$E_{da} = \left\{ \begin{array}{l} \left(a, do\right) \in A \times DO \mid \exists \alpha \in A_{\alpha}, \\ \alpha = \left(kl_i, sm_i, st_{sm_i}, hsm^{-1}, LT\right) \end{array} \right\}$$

– сторони графа.

3. СЕРВІСНА АРХІТЕКТУРА ЕБ

Як було сказано раніше при об'єднанні електронних бібліотек виникають проблеми з відмінностями в архітектурі ПЗ. Ми вважаємо, що використання сервіс-орієнтованої архітектури є ключовим елементом досягнення інтероперабельності. У такому середовищі складне програмне забезпечення може бути спроектоване на основі сервісів, що є у розпорядженні і які доступні не лише локально, але і через Інтернет. У такому контексті для побудови електронних бібліотек ми пропонуємо використовувати архітектуру сервіс-орієнтованої електронної бібліотеки (Service-Oriented Digital Library architecture – SODL) [17], це забезпечить зручний спосіб для досягнення побудови колективної пам'яті.

Для забезпечення вільного доступу служб до бібліотеки вона має бути реалізована на відкритій архітектурі. Тобто взаємодія служб має бути описана за допомогою стандартизованих правил і середовища, що дасть можливість вільно вводити новий сервісний елемент, не перебудовуючи систему наново.

Для того, щоб забезпечити динамічне налаштування електронної бібліотеки, необхідно забезпечити механізми, які виконуватимуть цю функцію. Таким механізмом є введення семантики в середовище функціонування веб-сервісів. Цей підхід отримав назву Semantic Web, і стосується не тільки семантичного опису сервісів а також і сематичного опису інформаційних ресурсів Інтернету (веб-сторінок, документів тощо.)

Під сервісом ми розуміємо деякою послугу мета якої задовольнити запити користувача. Сервіси реалізуються за допомогою веб-сервісів. Веб-сервіс це програмне застосування. Основна відмінність між сервісом і веб-сервісом полягає в тому, що той самий сервіс одночасно можна реалізувати різними веб-сервісами. У свою чергу веб-сервиси можуть між собою об'єднуватися, створюючи новий веб-сервіс, цей процес називають композиція веб-сервісів.

Сервіси електронної бібліотеки можна класифікувати за різними аспектами. Зокрема результатом виконання деякого сервісу на вимогу запитуючої сторони буде результат у виді структурованої інформації про електронні об'єкти які містяться в ЕБ. Тому якщо розглядати ЕБ як замкнене середовище, очевидно, що інформаційне наповнення ЕБ не зазнало змін, сервіси з такою властивістю будемо називати сенсорні сервіси.

Поняття сенсорних сервісів для ЕБ ми ввели тому, що ці сервіси аналогічно до сенсорів фільтрують інформаційне середовище, фактично не впливаючи на зміст цього середовища.

В рамках підходу Semantic Sensor Web [18] ми додатково анотуємо результати виконання сервісів. Тому на відміну від класичної побудови семантичного середовища для веб-сервісів, де результат не анотується, ми передбачаємо анотування метаданими результатів виконання веб-сервісів. Характер метаданих повинен передбачати історію, спосіб, локалізацію та час отримання результатів.

Для побудови електронної бібліотеки на основі сервіс-орієнтованого підходу необхідно виділити сервіси, на основі яких виконуватиметься композиція [19]. Це обумовлено тим, що завдання композиції простіше вирішувати і, крім того, вимоги до електронної бібліотеки є такими, що в першу чергу необхідно вирішувати композиційні завдання. Базові сервіси виділено в роботі [20].

Нехай ми маємо запит r, у якого є початкові параметри r_{in} і бажані параметри виходу r_{out}, необхідно знайти веб-сервіс w такий, щоб виконував r причому виконувалися умови:

$r_{in} \supseteq w_{in}$

$r_{out} \supseteq w_{out}$

Проблема пошуку веб-сервісу, який може самостійно виконати запит r носить назву дослідження веб-сервісів (Web Service Discovery – WSD). Коли за допомогою одного веб-сервісу неможливо досягнути виконання запиту r, потрібно утворити композицію кількох веб-сервісів $\left\{w^1, w^2, w^3, ..., w^n\right\}$ послідовним чи паралельним способом, таким чином, щоб для всіх $w^i \in \left(w^1, w^2, w^3, ..., w^n\right)$, причому w^i_{in} може бути заснований на w^i_{out}, та виконувалось відношення $\left(r_{in} \cup w^1_{out} \cup ... \cup w^n_{out}\right) \supseteq r_{out}$. Ця проблема носить назву композиції веб-сервісів (Web Service Composition – WSC).

Більш складною проблемою є проблема

композиції веб-сервісів.

В процесі функціонування сервіс-орієнтованої електронної бібліотеки середовище в якому функціонують веб-сервіси постійно змінюється. Зміна середовища породжується двома основними факторами:

по-перше проблеми, які породжують розподілені системи часові затримки і ненадійність транспортного протоколу, недостача пам'яті спільного використання між частинами розподіленої системи, проблеми відмови доступу та паралельних запитів, а також проблеми пов'язані з програмною несумісністю в наслідок оновлення частини розподіленої системи;

по-друге людський фактор, в процесі обслуговування користувача, можуть змінитися вимоги користувача, а отже це певним чином впливатиме на результат. Пересічні користувачі цільової аудиторії не мають чіткого уявлення про архітектуру бібліотеки, а отже не можуть на перед чітко визначитися з цілями які повинні задовольняти сервіси.

Ми будемо розглядати композицію на основі ціле-орієнтованої парадигми, тобто виходячи початкових умов та наявної множини сервісів здійснити композицію. Причому, оскільки веб-сервіси розміщені в семантичному середовищі, то вибирати такі плани композиції, які можуть бути корисними для кінцевого користувача. Тобто на відміну від класичної постановки задачі, від специфікації цілі до пошуку сервісів, які зможуть цю ціль досягти, виходити з того припущення, що наявна множина сервісів може досягнути деякі наперед невідомі цілі, які можуть бути обрані користувачем.

В такій складній системі цілі, які постають перед композицією можуть змінюватися. Ми також виходимо з того, що знання про оточуючий світ є не повними а отже і цілі є неповними. Для таких конфліктних цілей Horst Rittel і Melvin Webber ввели поняття вікід (wicked) задачі. Тому класичні методи планування виявилися неефективними. В середовищі де цілі не є чітко визначеними і є динамічний процес планування приймає інше смислове значення, а саме, план виконання дій не є однозначним алгоритмом досягнення цілі, функція планів можна сформулювати наступним чином [21]: перевірка ресурсів; починають координаційні процеси або допомагають спростити координаційні процеси на початку; встановлюють відповідальність та індетифікацію; трековий прогрес і що більш важливо генерація намірів; впізнають та управляють ризиками; підтримують імпровізацію, і що саме важливо формалізують представлення людини (користувача) про проблему яку план намагається вирішити.

Така комбінація функціональних властивостей призводить до того, що плани необхідно додатково досліджувати і проводити репланування. Тобто змінювати плани під час виконання цих планів.

4. ПРАКТИЧНА РЕАЛІЗАЦІЯ

Реалізація викладених положень є надзвичайно складною, оскільки постає ряд задач прикладного характеру. На даний момент нема довершеного фреймворку для моделювання автоматичної композиції веб-сервісів. Однак на наш погляд найбільш перспективним є WSMT [22] та WSMX [23], які базується на WSMO.

WSMO[1] (WSMO – Web Service Modeling Ontology) підхід в якому відображено всі аспекти які пов'язані з семантичними сервісами, які доступні через веб інтерфейс. Кінцевою ціллю цього підходу являється представлення інформації для автоматичного машинного вирішення задач дослідження, вибору, композиції, співставлення, виконання та моніторингу.

WSMO оперує 4 головними поняттями:

Онтології представляють термінологію яку використовують інші компоненти WSMO.

Веб-сервіси представляють обчислювальні об'єкти, таким чином, що в деякій предметній області ці об'єкти являють собою сервіси.

Цілі представляють побажання користувача, відповідно для задоволення яких можна відшукати веб-сервіс.

Посередник – описує ті елементи, які відповідають за вирішення питань функціональної сумісності між іншими елементам WSMO.

Синтаксис WSMO визначається мовою WSML – Web Service Modeling Language.

Сервіс підключається до середовища **WSMT** за допомогою WSDL інтерфейсу. Тому нами було розроблено два сервіси для існуючого вільного ПЗ управління електронними бібліотеками Eprints[2]. Сервіс getEprint який повертає метадані для запису з вказаним на вході id. Другий сервіс є набагато потужнішим. Цей сервіс має назву searchEprint, він дозволяє виконувати повнотекстовий пошук та пошук в метаданих електронної бібліотеки. На сонові останнього сервісу можна моделювати ряд інших сервісів, таких наприклад як сенсорні сервіси класифікації. Кожен сервіс має інтерфейс який

[1] http://www.wsmo.org/

[2] http://www.eprints.org/

відповідає специфікації WSDL 1.1. Дані розробки викладені на офіційному сайті Eprints а також обидва сервіси впроваджено на сайті http://eprints.zu.edu.ua/.

5. ВИСНОВОК

В даній роботі коротко викладено теоретично-практичні основи створення ЕБ з використанням семантичних технологій. Зроблено огляд провідних Європейських проектів з інтеграції технології Semantic Web в ЕБ. В статті також вказано на принципи застосування Semantic Web до сервісів ЕБ.

Проте проблематика створення таких ЕБ потребує подальшого вивчення, зокрема в роботі не формалізовано поняття онтології. Водночас також не вказано і типи зв'язків які можуть бути між екземплярами та онтологією. Проте вже зараз можна стверджувати, що необхідно буде вирішувати проблему вирівнювання між онтологіями, яка виникне в наслідок інтеграції двох або більше семантичних ЕБ.

6. СПИСОК ЛІТЕРАТУРИ

[1] Licklider, J. C. R. Libraries of the Future. Cambridge : MIT Press, 1965.

[2] *Going digital: a look at assumptions underlying digital libraries.* Marshall, D. M. Levy and C. C. 8, 1995 p., Communications of the ACM, T. 38, Pp. 77-84.

[3] *What are digital libraries? competing visions.* Borgman, C. L. 3, 1999 p., Information Processing and Management, T. 35, Pp. 227-243.

[4] IEEE Technical Committee on Digital Libraries. http://www.ieee-tcdl.org.

[5] Neuhold, Erich J. Position Statement Past Chairman. http://www.ieee-tcdl.org/posstatement.html.

[6] Ismail Fahmi, Junte Zhang, Henk Ellermann, Gosse Bouma. SWHi System Description: A Case Study in Information Retrieval, Inference, and Visualization in the Semantic Web. *The Semantic Web: Research and Applications, 4th European Semantic Web Conference.* Innsbruck, Austria : Springer, 2007, Pp. 769-778.

[7] *Porting Cultural Repositories to the Semantic Web.* Omelayenko, B. Tenerife, Spain : 2008. Proceedings of the First Workshop on Semantic Interoperability in the European Digital Library (SIEDL-2008). Pp. 14-25.

[8] Wielemaker, J., Hildebrand, M., Ossenbruggen, J.R. Van. Using Prolog as the fundament for applications on the semantic web (2008). *Proceedings of the 2nd Workshop on Applicatiions of Logic Programming and to the web, Semantic Web and Semantic Web Services.* Porto, Portugal : 2007.

[9] *SeRQL: A Second Generation RDF Query Language.* Broekstra J, Kampman A. Proc SWAD-Europe Workshop on Semantic Web Storage and Retrieval 2003.

[10] Eric Prud'hommeaux Andy Seaborne. SPARQL Query Language for RDF. *W3C.* 2008 p. http://www.w3.org/TR/rdf-sparql-query.

[11] *Solutions for a Semantic Web-Based Digital Library Application.* Martínez, Andrew Russell Green and José Antonio Villarreal. 2008. First Workshop on Semantic Interoperability in the European Digital Library.

[12] *SPARQLeR: Extended Sparql for Semantic Association Discovery.* Krys Kochut, Maciej Janik. 2007. 4th European Semantic Web Conference (ESWC2007). http://www.eswc2007.org/pdf/eswc07-kochut.pdf.

[13] *Semantic Maps and Digital Islands: Semantic Web technologies for the future of Cultural Heritage Digital Libraries.* A. Felicetti, H. Mara. Tenerife, Spain : 2008. SIEDL 2008: Semantic Interoperability in the European Digital Library. Pp. 51-62.

[14] RDF Core Working Group. Resource Description Framework (RDF). *Resource Description Framework.* W3C. http://www.w3.org/RDF/.

[15] *Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries.* Marcos André Gonçalves, Edward A. Fox, Layne T. Watson, Neill A. Kipp. 2, ACM New York, NY, USA, 2004 p., ACM Transactions on Information Systems (TOIS), T. 22. ISSN:1046-8188.

[16] *A formal model of annotations of digital content.* Maristella Agosti, Nicola Ferro. 1, 2007 : ACM New York, NY, USA, T. 26.

[17] *A service-oriented architecture for digital libraries.* Yves Petinot, C. Lee Giles, Vivek Bhatnagar, Pradeep B. Teregowda, Hui Han, Isaac Councill. ACM New York, NY, USA, 2004. Proceedings of the 2nd international conference on Service oriented computing. ISBN:1-58113-871-7.

[18] *Web, Semantic Sensor.* Sheth, A., Henson, C. та Sahoo, S.S. 4, 2008 p., Internet Computing, IEEE, T. 12. 10.1109/MIC.2008.87.

[19] *A service-oriented architecture for digital libraries.* Yves Petinot, C. Lee Giles, Vivek Bhatnagar, Pradeep B. Teregowda, Hui Han, Isaac Councill. ACM New York, NY, USA, 2004. Proceedings of the 2nd international

conference on Service oriented computing.
[20] Інноваційні підходи до створення колективної пам'яті на основі парадигми Semantic Web. О.В., Новицький. Львів : ПП Вежа і Ко, 2008. Computer Science and Information Technologies 2008.
[21] Flexecution as a Paradigm for Replanning, Part 1. Klein, Gary. 5, Los Alamitos, CA, USA : 2007 p., IEEE Intelligent Systems, T. 22, Pp. 79-83. 1541-1672.
[22] Web Service Modeling Toolkit (WSMT). http://sourceforge.net/projects/wsmt/.
[23] WSMX (Web Service Modelling eXecution environment). http://www.wsmx.org/.
[24] OpenAcademia. www.openacademia.org.

Олександр Новицький, *молодший науковий співробітник Інституту програмних систем НАН України.*

Наукові інтереси: технологія Semantic Web та її прикладне застосування в електронних бібліотеках.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

PARADIGM SEMANTIC WEB IN THE CONTEXT OF DIGITAL LIBRARY: SERVICES AND INFORMATION INTEGRATION

O. Novytskyi

Institute of Software System NAS of Ukraine 40 Academician Glushkov Ave., 03187, Kyiv, Ukraine
alex@zu.edu.ua

Abstract: *This paper presents a several of theoretical ideas and applied technology that can be embodied in the creation of Semantic Digital Library (SDL). In particular, considerable attention paid to how the Semantic Web technology is used in various aspects of DL. A basic level in the DL that may be carriers of semantic description. Described the advantages of this approach. It also highlights issues that arise during the integration of classical electronic libraries and Semantic Web a place in these processes. Made a brief overview of the world's leading projects to create DL using Semantic Web.*

Keywords: *Semantic Web, digital library, WSMO, semantic annotation.*

INTRODUCTION

Electronic Library is understood from position the collective memory and a complex structure. Committee of IEEE Technical Committee on Digital Libraries, interprets this concept as a set of Digital libraries, digital museums, digital archives [1].

Collective memory forms the so-called knowledge portals and content, with a network of distributed resources.

The development of a collective memory requires the development of other areas such as: storage, user interface, classification, information search, management and conservation.

In addition to above discussed areas should also increase the degree of detail descriptive information. [2].

1. SEMANTIC WEB ISSUES IN THE CONTEXT OF DL

Using semantic technologies in digital libraries has been given attention in many projects such as: SWHi [3], eCulture [4], IPISAR [5], EPOCH та AMA [6].

The first problem is that each directory has its own search engine and uses a grammar to describe the metadata and data, including that it will never work on other systems. Out of this situation will be the presentation of information in machine understandable form, using RDF [7].

The second problem concerns directly to information: a huge variety of formats that are used for indexing data, is a major obstacle to integration.

To create a single conceptual layer, semantic information must be taken from the database, HTML-page, descriptive text, and metadata to be presented in a standard format in order to obtain the conceptual content of information created conceptual mapping.

One way to address this problem can be generalized formal model of annotations.

Another problem that arises when handling large amounts of heterogeneous information is to ensure appropriate services.

2. FORMAL MODEL OF ANNOTATIONS

The main two components of the electronic library is its content, and set the software to work with this content. To start, consider the content of EB. Information in libraries is described in terms of electronic facilities (Digital objects DO), which are multimedia content and metadata [8]. Formal model that suggested in [9], identifies two approaches to understanding the annotation: annotations as metadata or annotation as content.

Among the many models, we use a model that suggested in [8], [9] with some changes and refinements, which include the use of ontology. Let LT the set of types of links.

$H(k)$ set of identifiers of digital objects in time k.

Set of segments ST, we indicate that

$\forall st_{sm_i} \in ST,\ i = 1,2,...,n.$

SM defines the set of streams, so that $\forall sm_i \in SM,\ i = 1,2,...,n.$

Annotation can be viewed as the process enlargement of ontology O_i.

Each class of ontology O_i shall designate kl_i, a set of classes KL.

Annotation $a \in A(k)$ is tuple:

$$a = \begin{pmatrix} h_a \in H(k), \\ A_a \subseteq KL(k) \times LT \times ST(k) \times \\ \times SM(k-1) \times H(k-1) \end{pmatrix}$$

where h_a – own unique identifier annotations a, that is $h(h_a) = a$;

A_a set n-arity relations annotation a and is defined as the product sets KL, LT, ST, SM and H.

In the case of annotation of web documents, the formal model of change. Let A set all annotation a, and D set DO, accordingly, $DO = D \cup A$, a subset of the set DO to mark do, that is $do \in DO.$

Annotation Web-document called tagging graph:

$$G := \left(\left(DO, E_{da} \subseteq A \times DO\right)\right),$$

where $DO = D \cup A$ vertex graph;

$$E_{da} = \begin{Bmatrix} (a, do) \in A \times DO \mid \exists \alpha \in A_a, \\ \alpha = \left(kl_i, sm_i, st_{sm_i}, hsm^{-1}, LT\right) \end{Bmatrix}$$

– side of the graph.

3. SERVICE-ORIENTED DIGITAL LIBRARY ARCHITECTURE

We believe that the use of service-oriented architecture is a key element in achieving interoperabelnosti. To build digital libraries, we propose to use the architecture of service-oriented e-Library (Service-Oriented Digital Library architecture – SODL) [10], it provides a convenient way to achieve the construction of collective memory

To ensure a dynamic setting digital library must provide mechanisms that perform this function. This mechanism is the introduction of semantics in the operation environment of web services.

Digital library services can be classified according to various aspects. Services are in the process of change is not content DL, is called sensory services.

As part of the approach Semantic Sensor Web [11] we additionally anotuyemo results of services. Therefore, unlike classical building environment for semantic web services, which result not anotuyetsya, we anticipate the results of the metadata annotation of web services.

Basic services allocated in the [12].

We will consider the composition based on goal-oriented paradigm is based initial conditions and the existing set of services to make the composition. And, because Web services are located in the semantic environment, then choose the plan of composition that may be useful for the end user. That is, unlike the classic statement of the problem of specification aims to find services that can achieve this goal, go with the assumption that the available set of services can reach some pre unknown targets that may be selected by user.

This combination of functional properties has meant that the plans need further study and conduct replanning. That is changing plans during the implementation of these plans.

4. CONCLUSIONS

However, in our opinion the most promising compositions for framework Web-services is WSMT [13] and WSMX [14], which is based on WSMO.

We developed two services for the existing free software management digital libraries Eprints1: service getEprint and searchEprint. Each services the interface that meets the specification WSDL 1.1. What allows these services to connect WSMT.

This paper briefly outlined the theoretical and practical bases of DL using semantic technologies. An overview of European projects to integrate technology in the Semantic Web DL. The article also stated on the application of principles of Semantic Web services to the DL.

However, creating such problems DL needs further study, particularly in the no formal notion of ontology. At the same time also has both types of connections that can be between instances and ontologies. But now we can say that it will be necessary to solve the problem of alignment between ontologies, which arise as a consequence of the integration of two or more semantic DL.

5. REFERENCES

[1] IEEE Technical Committee on Digital Libraries. [Online] http://www.ieee-

[1] http://www.eprints.org/, http://files.eprints.org/401/

tcdl.org.
[2] Neuhold, Erich J. Position Statement Past Chairman. [Online] http://www.ieee-tcdl.org/posstatement.html.
[3] Ismail Fahmi, Junte Zhang, Henk Ellermann, Gosse Bouma. SWHi System Description: A Case Study in Information Retrieval, Inference, and Visualization in the Semantic Web. *The Semantic Web: Research and Applications, 4th European Semantic Web Conference.* Innsbruck, Austria : Springer, 2007, p. 769-778.
[4] *Porting Cultural Repositories to the Semantic Web.* Omelayenko, B. Tenerife, Spain : s.n., 2008. Proceedings of the First Workshop on Semantic Interoperability in the European Digital Library (SIEDL-2008). p. 14-25.
[5] *Solutions for a Semantic Web-Based Digital Library Application.* Martínez, Andrew Russell Green and José Antonio Villarreal. 2008. First Workshop on Semantic Interoperability in the European Digital Library.
[6] *Semantic Maps and Digital Islands: Semantic Web technologies for the future of Cultural Heritage Digital Libraries.* A. Felicetti, H. Mara. Tenerife, Spain : s.n., 2008. SIEDL 2008: Semantic Interoperability in the European Digital Library. p. 51-62.
[7] RDF Core Working Group. Resource Description Framework (RDF). *Resource Description Framework.* [Online] W3C. http://www.w3.org/RDF/.
[8] *Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries.* Marcos André Gonçalves, Edward A. Fox, Layne T. Watson, Neill A. Kipp. 2, s.l. : ACM New York, NY, USA, 2004, ACM Transactions on Information Systems (TOIS), Vol. 22. ISSN:1046-8188.
[9] *A formal model of annotations of digital content.* Maristella Agosti, Nicola Ferro. 1, 2007 : ACM New York, NY, USA, Vol. 26.
[10] *A service-oriented architecture for digital libraries.* Yves Petinot, C. Lee Giles, Vivek Bhatnagar, Pradeep B. Teregowda, Hui Han, Isaac Councill. s.l. : ACM New York, NY, USA, 2004. Proceedings of the 2nd international conference on Service oriented computing. ISBN:1-58113-871-7.
[11] *Web, Semantic Sensor.* Sheth, A., Henson, C. e Sahoo, S.S. 4, 2008, Internet Computing, IEEE, Vol. 12. 10.1109/MIC.2008.87.
[12] *Innovation Approaches to design of collective memory on the paradigm base of Semantic Web.* O.V. Novitsky. Lviv : PP Vezha and Co, 2008. Computer Science and Information Technologies 2008.
[13] 13. Web Service Modeling Toolkit (WSMT). [Online] http://sourceforge.net/projects/wsmt/.
[14] WSMX (Web Service Modelling eXecution environment). [Online] http://www.wsmx.org/.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

IMAGE STRUCTURE ANALYSIS BY 3-STAGES CLUSTERING

Roman Melnyk [1], Ruslan Tushnytskyy [2]

[1] Professor, Software Department, Lviv Polytechnic National University, 12, S. Bandery str., Lviv, 79013, ramelnyk@polynet.lviv.ua
[2] Post-graduate student, Software Department, Lviv Polytechnic National University, 12, S. Bandery str., Lviv, 79013, ruslan.tushnytskyy@gmail.com

Resume: *An approach for decomposition of visual images by clustering and breaking them down into geometric figures is considered. Multilevel hierarchical clustering algorithm to form three emphasized levels of clusters such as rectangles, closed regions and integrated areas is proposed. Advantages of such decomposition in three stages are as follows: images covered by rectangles are planned to be formatted and compressed, image fragments could be taken for the preliminary pattern recognition or could easily be corrected, hierarchically constructed fragments are good material to form pattern features for searching procedures. The algorithm complexity, the proposed approach of scanning searching area to reduce it, the rolling up criteria and key parameters for its control are investigated. The results of pattern analysis by structure features for some practical problems are presented in the article.*

Keywords: *cluster, clustering, visual pattern, hierarchical tree, rolling-up algorithm, scanning area, rectangles, integrated areas, structured images, structural features.*

1. INTRODUCTION

The clustering methods are widespread for visual patterns segmentation, data mining, indexing and searching. [1−9]. In particular, the work [1] contains the classification of the clustering methods and the approach for contouring chosen clusters. The works [2], [3] describe graph's models for images and clustering algorithms for its segmentation. The works [4], [5] are devoted to the clustering approach for experimental data covered by a grid with predetermined step. Hierarchical methods were used for decomposition and enhancement of medical images and fingerprints [6] − [7]. Various approaches are used for improving and decomposition of images, among them being assessment of image fragment local characteristics [8], forming regions of special figures [6] or clustering according to criteria of statistical dependencies [7]. These works, as well as a number of others, do not consider the variety of control parameters, by which researcher could investigate different dependences of clustering results, e.g., such as brightness threshold and intensity sensitivity for number of clusters to be formed. As a rule they are based on one stage clustering procedures.

Different agglomerative methods differ by the way that similarity between two clusters is updated. If the linkage satisfies a cluster aggregative inequality then the algorithm can be implemented efficiently at a time complexity of $O(N^2)$ [10]. A new recent method [11] uses a sophisticated merging criterion which takes the cluster internal structure into account. The multilevel hierarchical clustering algorithm in [12] are mainly planned for data analysis, has one stage button-up procedure and proposes new linkage criteria to reduce the computation complexity from $O(N^3)$ to $O(N^2)$.

Standard agglomerative clustering methods use ordinary similarity distances between clusters to be merged. The proposed approach is based on a number of merging criterion: value of common border between clusters, difference between their weighted brightness, difference between square values, etc. Moreover, merging is being performed if constraints allow it. The work is based on three stages clustering algorithm: covering an image by rectangles, closed regions and integrated areas Constraints are formed to control form of rectangles or brightness of closed regions and integrated areas.

Some investigations were held to describe different dependences: number of clusters, parameters of closed regions and integrated areas due to criteria for merging: cluster brightness threshold, algorithm sensitivity or shape and dimension constraints.

The full procedure allows to formalize visual image description by breaking them down into integrated areas, closed regions or rectangles [Fig.1].

An access to output fragments of every stage allows them to be corrected and broken into fragments. We would like to notice that we didn't find such image fragments classification in the referenced papers and others read. Additionally we note that *Microsoft Visual Studio.NET* disposes a class Region to paint figures with non regular shape.

When applying three stages clustering algorithm to practical images algorithmic complexity and computation time of image decomposition process would be taken into account. For this reason, scanning searching procedures on different stages of the algorithm are developed.

For large database with over tens of thousands of visual patterns, effective indexing becomes an important issue the content-based image retrieval systems (CBIR). The existent universal CBIR systems attribute to one of three categories depending on approach of extracting features, for example color [15 − 16], shape [17 − 18], structure and location [19 − 20]. Presented 3-stages clustering algorithm extracts some pattern structural properties (pattern features) which were taken to analyze and classify images.

2. TASK OF IMAGE DECOMPOSITION

The tasks of the decomposition algorithm when being applied to visual patterns are to solve the following problems: 1) to get full description for all parts of image with its all possible characteristics; 2) to ensure an access for every part of pattern independently of its position on three levels hierarchical tree: integrated areas → closed regions → rectangles. Such access is planned to be used in searching procedures for pattern and its fragment recognition.

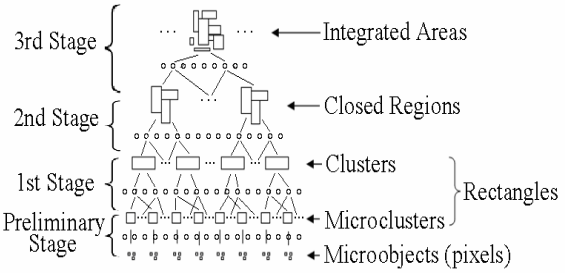


Fig. 1 – Fragment of multilevel hierarchical tree

To solve these complex tasks the multilevel three stages clustering algorithm based on traditional binary hierarchical tree is developed. Three stages of the process, the nodes of interest and a small fragment of the multilevel tree are shown on Fig. 1. Three stages are emphasized because they have different searching criteria to build different geometric figures. For user it is more interesting to consider three levels tree extracted from full tree presented on Fig. 1 with earlier defined pattern fragments (areas, regions and rectangles) as nodes.

Before the clustering algorithm begins its work the preliminary procedure to prepare leafs of the hierarchical tree has to start. Non divisible microobjects, i.e., pixels of digital image must be united to form microclusters − smallest non divisible rectangles. The microclusters are fully or partly filled [Fig. 2(a)] in the case of black-and-white image or are combined from pixels of different brightness for grey image [Fig. 2 (b)].

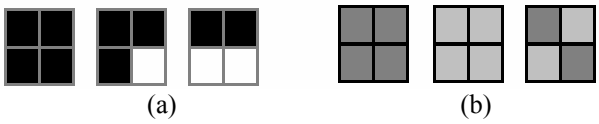


Fig.2 – Black-and-white and grey clusters

On the preliminary stage we must to solve the problem:

It is necessary to obtain a partitioning $\bar{O}$ of the set of microobjects $O=\{o_1, o_2, \dots, o_N\}$:

$$\left.\begin{array}{l} \overline{O} = \{O_1, O_2, ..., O_n\}, \\ n \to \min, \end{array}\right\} \quad (1)$$

so that a total number of partitions is minimized while satisfying the constraints:

$$F_k^- \le F_k(X_i) \le F_k^+, \quad (k = \overline{1,m}; i = \overline{1,n}), \quad (2)$$

where n is number of microclusters $O_1, O_2, \dots, O_n$; F_k^-, F_k^+ − are boundary values of the k-th feature function. The functions $F_1, F_2, \dots, F_n$ define coordinates, brightness, dimension, number of pixels, shape, color, ratio between filled and empty cells etc.

The cluster features or parts of them are also assigned for clusters on the higher levels of the hierarchical tree. So to make pattern decomposition by rectangles, closed regions and integrated areas we must to solve the problem (1) three times. We solve the optimization problem by the rolling-up procedure, i.e., algorithm of bottom-up merging of clusters to get minimal number of segments covering a pattern. The stages differ between themselves by merging criterion and constraints. The rolling-up procedure divides a set of objects into groups having the same or similar characteristics. Tree nodes not satisfying the constraints and subsequently not being the subjects for further rolling-up process are transferred to the result set of clusters.

The visual pattern area is being covered by grid with steps selected from a set of values: 1×1, 1×2, 2×1, 2×2 etc. Some scanning procedures to solve the

problem (1) on the preliminary stage of the algorithm and to build microclusters as rectangles sized by selected steps of grid are developed. A cluster number on higher stages depends from a microcluster number and its features. That is why we developed alternative algorithms for scanning procedures on the preliminary stage of the algorithm. The trivial approach to get microclusters is to describe a grid cell as microcluster, i.e., to define for it all features and to assign them to it. For reason of a lack of paper place other preliminary scanning algorithms are not considered in the paper.

3. CONTROL PARAMETERS

For grey image every microobject (pixel), is characterized by brightness value ranging from black to white. Brightness b of microobject we define by equation expressing its relative filling as percentage of black color:

$$b = (256 - c_i) \times 100 / 256, \qquad (3)$$

or as percentage of white color:

$$b = c_i \times 100 / 256 \qquad (4)$$

where c_i (i = 1,2,3) is a value of one from the R (G or B) components of grey pixel.

Fig. 2(b) shows rectangles by different brightness values.

The equations (3), (4) are applicable only for microobjects. To denote integral microcluster brightness we use B_{rk}. We define it as a mean weighted value for two microobjects (microclusters) indexed by i and j and merging to a microcluster numbered by k:

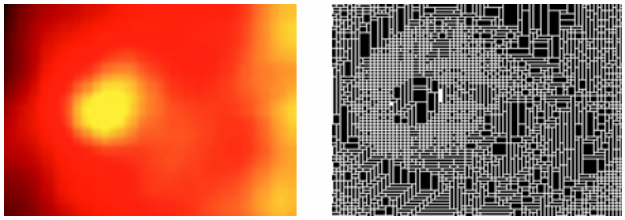
$$B_{rk} = \frac{b_i \times a_{i1} \times a_{i2} + b_j \times a_{j1} \times a_{j2}}{a_{i1} \times a_{i2} + a_{j1} \times a_{j2}}, \qquad (5)$$

where b_i, b_j are the brightness values, a_{i1}, a_{i2} a_{j1}, a_{j2} are values of microobject geometric sizes.

The equations (1) – (5) are being used on the preliminary and 1st stages of the algorithm where microcluster and cluster as rectangles are considered. Fragment of medical pattern clustered by rectangles is illustrated by Fig. 3.

To denote integral closed region and integrated area brightness we use B_{Rk} and B_{Ak}. We define them as a mean value of cluster or closed region brightness of all n objects creating closed region or integrated area:

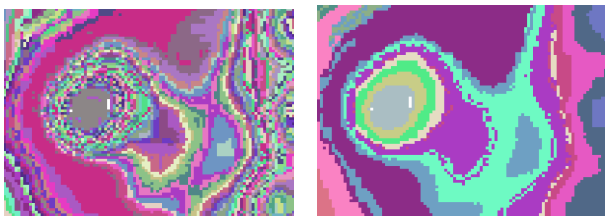
$$B_{Rk} = 1/n \times \sum b_i; \qquad B_{Ak} = 1/n \times \sum B_{R\,i} \qquad (6)$$



(a) **(b)**

Fig.3 – Clustered pattern. (a) Original. (b) Pattern after preliminary and 1st stages of the algorithm

The equations (6) are being used on the 2nd and 3rd stages of the algorithm where closed regions and integrated areas are considered. Fragment of medical pattern clustered by regions and areas are illustrated by Fig. 4.



(a) **(b)**

Fig.4 – Clustered pattern after 2nd (a) and 3rd (b) stages of the algorithm (colors are random)

4. ALGORITHM ACCURACY

For black-and-white image there are two types of microcluster (rectangle): completely black and with a number of white positions.

Ratio between a number of black pixels n_{bi} entering i^{th} microcluster and its total number of pixels n_{ti} we use as i^{th} cluster brightness: $r_{bi} = n_{bi} / n_{ti}$. For example, Fig. 2(a) shows for microcluster brightness values 1, ¾, ½.

We note: S_P – area of pattern (number of black pixels), S_{oi} – area of cluster, S_{Pc} – area of the pattern got as coverage by rectangles: $S_{Pc}=\sum S_{oi}$, where i runs over full set of clusters.

The brightness parameter of rectangle we use to control accurate or approximation clustering results. For accurate clustering we have $S_{Pc} = S_P$.

In practice we have to reduce extra or noise information and then it is useful to consider the case $S_{Pc} < S_P$. We call this procedure cleaning or reduction. This process is similar to filtration of thin effects in visual patterns.

Opposite case is if $S_{Pc} > S_P$. This procedure is named as renewal or addition. We mean here, that by a noise in the pattern some needed pixels are destroyed or lost and its picture does not correspond to generally accepted one.

The brightness parameter is a key to control ratio between S_{Pc} and S_P, i.e., to control: 1) a numbers of

microclusters and clusters not taking part in merging of the rolling-up process; 2) a number of white pixels included in the clusters covering the pattern; 3) numbers of clusters and closed regions got by pattern clustering.

We consider accuracy for image covered by clusters or covered by integrated areas:

$$E_C = S_{Pc} / S_P, \qquad E_{IA} = IA_b / IA_{100}. \qquad (7)$$

where IA_b – number of integrated areas, got by clustering with user defined brightness parameter b, IA_{100} – number of integrated areas, got by clustering with brightness parameter b = 100%.

5. DECOMPOSITION ALGORITHM

To solve the problem (1) we consider the rolling-up algorithm having input objects as microclusters, clusters and closed regions and output objects as clusters, closed regions and integrated areas. To use one description for all these object we call them as Clusters. We denote them by Q_1, Q_2, …. The set X_i belongs to the i-th level of a tree and contains Clusters. On the *i-th* step of the algorithm merging Clusters create new Clusters by which the set X_{i+1} is being filled.

The sequence of steps in the ***rolling-up algorithm*** is as follows:

0. Prepare initial data: i=1, the set X_i is being filled by input Clusters; $C_i = X_i$, $B_i = \varnothing$.

1. Form for every $Q_k \in X_i$ a set of touching neighbors within the grid by searching procedure.

2. Calculate the criterion function values for all pairs of adjacent Clusters.

3. For every two candidates to be merged and satisfying all constraints create the list $L(F_{ki})$ of criterion values by decreasing order.

3. Delete from $L(F_{ki})$ pairs having intersected indexes beginning to search them from bottom.

4. Merge first pair of Clusters from the list and put new Cluster to the set B_i, correct the set C_i, form the new set X_{i+1} of Clusters.

5. Increasing i=i+1 repeat steps 1–4 until merging takes place.

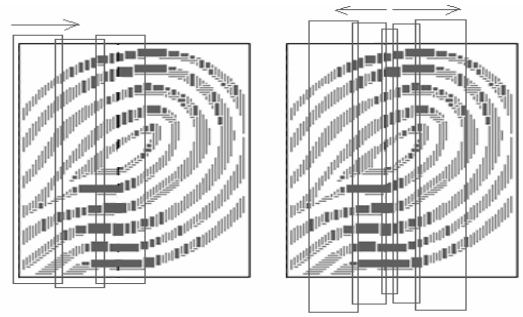
The last set X_L contains output Clusters

6.ALGORITHM COMPLEXITY OPTIMIZATION

Computation complexity of the clustering algorithm could be determined by a number of criterion function calculation, comparison and merging number. In the worst case, when merging number is 2N-1 (N is full number of input Clusters) complexity of clustering algorithm exceeds $O(N^2)$. Theoretically to calculate all functions (needed and extra) it does cost $O(N^3)$. The main operations of the algorithm are related with searching of candidates for merging.

To reduce the algorithm complexity we could mainly at the stage of searching Cluster candidates to be merged. The searching area for the rolling-up procedure we plan as a part of full pattern space. This part is as a rectangle moving through the pattern rectangle by alternative rules. It scans the whole surface of image [see Fig. 7]. Area limits for searching Clusters – candidates to be merged – are put on one axe of coordinates. Decreasing input data volume we reduce the algorithm complexity from value $O(N^2)$ to a set of smaller values $O(n_i^2)$, because the number n_i of Clusters in scanning area is smaller than the number N in full pattern area, i.e., n_i<N. We would like to mention that searching procedure repeats its work step by step covering the whole pattern area by a sequence of smaller scanning areas.

All Clusters of two coordinates pattern space from Fig. 7 are stored in linear data structure containing description of the Cluster model and a pointer for neighboring Cluster [Fig. 8].



(a) **(b)**

Fig.5 – Scanning area of the algorithm

First rule for scanning is illustrated by Fig. 5(a). The dashed-lined rectangle defines the scanning working area restricted by the right border. Scanning to search for neighboring Clusters is from left to right within the rectangle. After some searching and merging steps the rectangle is being moved to right.

Fig. 5(b) illustrates the second rule of scanning. The rectangle is being restricted both by the right and the left borders. An image is being scanned from left to right within the rectangle. To search for neighboring Clusters the rectangle itself has two directions for moving: to left and to right. Cluster for which neighboring Cluster are being searched we accept as the center of the scanning rectangle.

The scanning area has adaptive property: dimension of scanning rectangle could be of image size or could be formed dynamically depending from Cluster size. For example, at the beginning the

scanning area size is being initialized by layout grid steps and further they are being increased step by step. This adaptive property essentially reduces for current Cluster a number of neighbors to be compared.

The experiments presented in the Table 1 confirmed efficiency of scanning area: minimal time economy is by one order and computation advantage is better for bigger images.

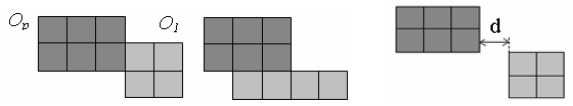
Table 1. Clustering algorithm without and with scanning area

Image dimension	Cluster number	Time without scanning area, s	Time with scanning area, s
120x178	5 155	0,751	0,07
149x150	7 311	0,661	0,05
175x216	8 210	0,450	0,05
141x169	11 952	1,112	0,07
156x173	13 617	2,193	0,09
171x213	14 136	3,074	0,140

7. CLUSTERING TO CLOSED REGION AND INTEGRATED AREAS

By the 2nd and 3rd stages of the algorithm we solve the problem (1) having rectangles $Q(Q_1, Q_2, Q_3, \ldots, Q_N)$ and closed region*s* $R(R_1, R_2, R_3, \ldots, R_N)$ instead of set O. A task for the rolling-up procedure is to find partitioning $\check{R}=\{R_1, R_2, \ldots, R_n\}$ and $\check{A}=\{A_1, A_2, \ldots, A_n\}$ instead of set $\bar{O}=\{O_1, O_2, \ldots, O_n\}$ so that a total number of partitions is minimized: $n\rightarrow\min$, while satisfying the constraints from (1).

Every set R_i contains rectangles included to one closed region by the rolling-up procedure. Criterion for rectangles to be merged is a number of common edges [Fig. 6(a)].



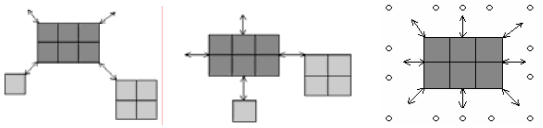
(a) **(b)**

Fig.6 – Merging criterion (a) and distance between clusters (b)

Constrains are formulated by feature functions $F_{i..}$ Second criterion for Cluster rectangle to participate in merging is minimal distance d in pixels between their closest rectangles [Fig. 6(b)]. Distance d is measure to be controlled by user.

We extended the distance criterion for "diagonal" and "main" distances [Fig. 7(a)].

In particular, diagonal distance is useful to process images having ring structure, e.g., blood cells, fingerprints etc.



(a) **(b)**

Fig.7 – Diagonal and main distances between clusters (a), a set of coordinates for a neighboring Cluster (b)

To realize the algorithm to solve the problem (1) with such objects as rectangles and closed regions we use new procedure to search candidates to be merged. The procedure is being used on the 1st step of the ***rolling-up algorithm***. The main its steps are as follows:

1. Select a current primary Cluster not belonging to any result Cluster.

2. Form a set of coordinates P around the Cluster from 1st step. The set includes all possible points, where neighboring Clusters could be located including the diagonal and main distances [Fig. 7(b)].

3. Look for a secondary Cluster to select it while satisfying the following conditions:

– it does not belong to any result Cluster;

– one or more coordinates of set P belong to the secondary Cluster.

4. Repeat step 3 until neighboring secondary Cluster exists.

5. Repeat all step until primary Cluster exists.

Having the candidates list the ***rolling-up algorithm*** continues its steps to build results Clusters (regions and areas) by hierarchical tree.

8. STRUCTURE FEATURES

As results the 3-stages clustering algorithm extracts from image some characteristics which we assign to the set of structural features. In particular: MC − number of microclusters, C, CR, IA − numbers of clusters, regions and areas from three stages. Also we get brightness parameters and square values for all objects.

We use these numbers and parameters to define structural properties by which one could estimate the structure degree of image, in particular:

1) structure coefficients corresponding to merging degree by one stage of the algorithm (an index specifies the algorithm stage):

$$K_s^1 = C / MC, \quad K_s^2 = CR / C, \quad K_s^3 = IA / CR \quad (8)$$

2) structure coefficients corresponding to merging degree by two stages of the algorithm:

$$K_s^{12} = CR / MC, \qquad K_s^{23} = IA / C. \qquad (9)$$

3) structure coefficients corresponding to

merging degree by three stages of the algorithm:

$$K_s^{123}(MC) = IA / MC. \quad (10)$$

To characterize inverse property (fuzziness) we use inverse values for (2-5), e.g.:

$$K_f^{123}(MC) = MC / IA. \quad (11)$$

The image structure coefficients we also assign to structural features. In some cases to widen the set we can in formulas (4-6) to use a total number of image pixels *PX* instead of *MC*:

$$K_s^{123}(PX) = IA / PX. \quad (12)$$

Each selected fragment i, in particular, area or region characterize the relative number of pixels:

$$vx_i(CR) = px_i(CR)/PX, \; vx_i(IA) = px_i(IA)/PX, \quad (13)$$

for which inequality is as follows:

$$\sum px_i(CR) \le PX, \quad \sum px_i(IA) \le PX, \quad (14)$$

where sum takes for all formed areas IA or region CR.

Some statistical properties could easily be considered as structural properties. For example:

1) Mean cluster, region or area dimension :

$$M(CR) = (1 / CR) \cdot \sum px_i(CR),$$
$$M(IA) = (1 / IA) \cdot \sum px_i(IA), \quad (15)$$

2) Cluster, region or area deviation

$$D(CR) = \sqrt{(1/CR) \cdot \sum (px_i(CR) - M(CR))^2},$$
$$D(IA) = \sqrt{(1/IA) \cdot \sum (px_i(IA) - M(IA))^2} \quad (16)$$

3) To estimate the image elementary object size/deviation features on different levels of decomposition we propose statistical relations:

$$R(CR) = \frac{M(CR)}{D(CR)}, \quad R(IA) = \frac{M(IA)}{D(IA)} \quad (17)$$

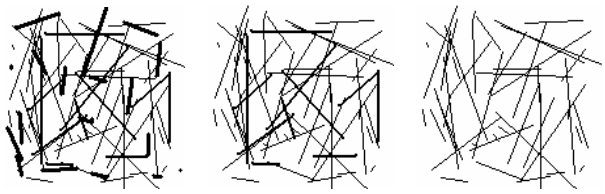
And for full image integral statistical relations

$$R_I(CR) = \frac{IA \cdot M(CR)}{D(CR)}, \quad R_I(IA) = \frac{IA \cdot M(IA)}{D(IA)} \quad (18)$$

All or part of above mentioned structural features are planned to be used for pattern and its fragments analysis and classification.

9. IMAGE STRUCTURE AND SURFACE ANALYSIS

The clustering algorithm was applied for material structure and surface analysis by using proposed structure features. Work [21] presents structured and non structured images investigation by the grid step and microcluster brightness parameter. Fig. 8 presents samples with lines by different width and length to imitate changing of structure complexity. Table 2 contains some structure coefficients and features illustrating structural changes especially by values of $K_s^{123}(MC)$ and $R_I(IA)$.



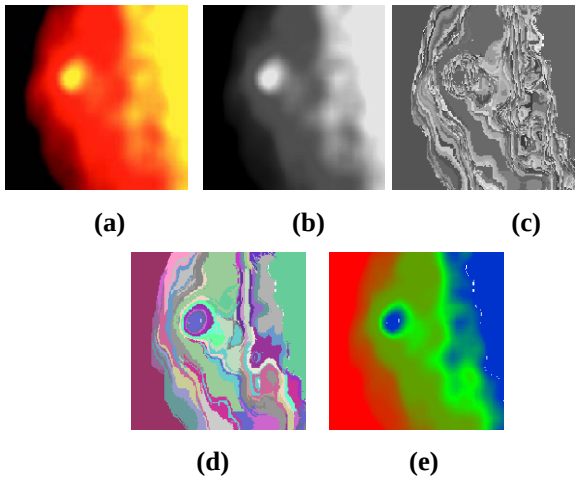
(a) (b) (c)

Fig. 8 – Images for structure analysis

Table 2. Structure coefficients and features

Structural characteristics	Visual pattern		
	a	b	c
$K_s^{123}(MC)$	0,00102	0,00214	0,00425
$R_I(IA)$	2,42940	2,87323	3,77030
MC	3912	2802	1882
C	1038	890	707
CR	11	11	13
IA	4	6	8

Fig.9 presents medical pattern structure analysis. The determined structural features of the mattern could be used for its indexing ans searchin in databases as well as for qualitative and quantitative estimation of the pattern given.



(a) (b) (c)

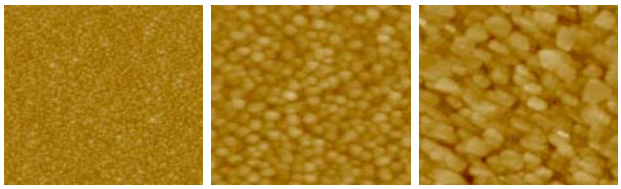
(d) (e)

Fig.9 – Medical pattern. (a) Input image. (b) Clusters. (c) Closed regions. (d) Integrated areas. (e) Closed regions coloured by their weighted brightness

The main structural coefficients and features are as follows: $K_s^{123}(MC)$ = 0,00182, MC = 36241, C = 25384, CR = 25272, IA = 66, $M(IA)$ = 549,106, $D(IA)$ = 1330,06.

By the experiment we can conclude that numbers of final Clusters (integrated areas) are of close values as it was expected.

The algorithm was applied for surface topology of freshly deposited surfaces [Fig.10].



(a) (b) (c)

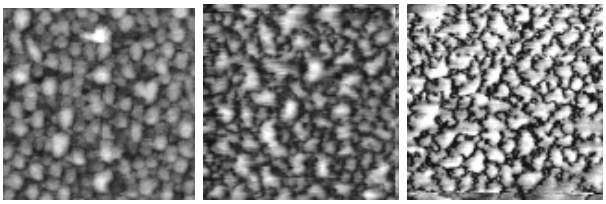
Fig. 10 – Surface topology of freshly deposited (a) Pt, (b) Au and (c) Ag surfaces. The average grain size is 3 nm for Pt, 8 nm for Au, 15 nm for Ag

By structural features from table 3 we can estimate the surface structure changes caused by different grain size, applied to polish the surface.

Table 3. Structure characteristics of metal surface

Structural characteristics	Visual pattern		
	a	b	c
$K_s^{123}(MC)$	0,00244	0,00294	0,00373
$R_I(IA)$	72,641	78,92	99,250
MC	21700	21700	21700
C	17498	15699	15449
CR	13973	11523	10895
IA	53	64	81
$M(IA)$	409,433	339,062	267,901
$D(IA)$	298,726	274,952	218,638

Close results were got by investigating the glass surfaces having different sizes of nanobubles [Fig.11]. Table 4 confirms: the more structure complexity the greater values have structural features.



(a) (b) (c)

Fig. 11 – A series images revealing nanobubbles on a glass surface

Fig 11(a) correspond to pH 9.4, Fig. 11(b) – pH 4.4, Fig. 11(c) – pH 5.6.

To widen a range of the algorithm application the specimen of polished Si surfaces were taken for investigation (fig12). Samples were elaborated by different grain sizes.

Table 4. Structure characteristics for glass surface

Structural characteristics	Visual pattern		
	a	b	c
$K_s^{123}(MC)$	0,00661	0,00754	0,00870
$R_I(IA)$	103,87	106,225	129,429
MC	11183	11134	9770
C	11112	10692	8836
CR	11072	10463	8076
IA	74	84	85
$M(IA)$	151,121	132,547	114,941
$D(IA)$	107,653	104,814	75,484

(a) (b) (c)

Fig. 12 – Surface images of a Si specimen following lapping with various abrasive types and sizes: (a) 14 – μm BC; (b) – 8 μm BC; (c) – 8 μm SiC

Results in table 5 and others which could be calculated confirmed that they are applicable for quality estimation of polished surfaces.

Table 5. Structure characteristics for Si surface

Structural characteristics	Visual pattern		
	a	b	c
$K_s^{123}(MC)$	0,00307	0,00388	0,00400
$R_I(IA)$	82,589	98,216	89,944
MC	22119	20079	22243
C	18834	17213	19226
CR	16112	14909	16746
IA	68	78	89
$M(IA)$	325,279	257,423	249,921
$D(IA)$	267,818	204,436	247,298

10. PROGRAM PACKAGE

The experimental program package with the user interface has been developed [21]. It controls all operation stages: image input, parameters, intermediate and output images, clustering options (full, specified), report of the algorithm work, result parameters for each clustering stage.

The program packages is programmed in C# language of *Microsoft Visual Studio 2008*.

11. CONCLUSION

The clustering algorithm for visual patterns analysis is developed and investigated. It is based on ideas: 1) classifications of pattern fragments as microclusters, clusters, closed regions and integrated

areas; 2) rolling-up algorithm to solve the problem of optimal coverage on three stages of the algorithm; 3) scanning area approach to reduce the algorithm complexity; 4) calculation of structure features to analyze material structure and surface.

The test results confirmed possibility to apply the program package for image analysis and processing to determine their properties and features, to find structural changes, to search defects in objects and materials. The package could be useful for specialists from medicine, metallurgy, spectroscopy and many others making conclusion by sample images.

All tests were performed by a computer with processor *AMD Athlon* 1.68 GHz, 768 Mb RAM.

12. REFERENCES

[1] A. Yip, C. Ding, T. Chan. Dynamic Cluster Formation Using Level Set Methods, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 28 (6) (2006). p. 877−889.

[2] L. Grady, E. Schwartz. Isoperimetric Graph partitioning for Image segmentation, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 28 (3) (2006). p. 469−475.

[3] M. Pavan, M. Pelillo. Dominant sets and Pairwise Clustering, *IEEE Trans. on Pattern Analysis and Machine Intelligence* 29(1) (2007). p. 167−172.

[4] Z. Yu, H.-S. Wong. GCA: A real-time grid based clustering algorithms for large data set, *Proc. of the 18th International Conference on Pattern Recognition (ICPR)* (2006). p. 740−743.

[5] K. Wilamowska, M. Manic. Unsupervised pattern clustering for data mining, *Proc. of the 27th Annual Conference of the IEEE Industrial Electronics Society (IECON)* (2001). p. 862–1867.

[6] S. Katz, A. Tal. Hierarchical mesh decomposition using fuzzy clustering and cuts, *ACM Transactions on Graphics* 22 (3) (2003). p. 954–961.

[7] R. Dosil, X.M. Pardo, X.R. Fdez-Vidal. Decomposition of three-dimensional medical images into visual patterns, *IEEE transactions on biomedical engineering* 52 (12) (2005). p. 2115−2121.

[8] L. Hong, Y. Wan, A. Jain. Fingerprint image enhancement: Algorithm and performance evaluation, *IEEE Transactions on Pattern Analysis and Machine Intellegence* 20 (8) (1998). p. 777−789.

[9] K. Anil, R. Dubes. Algorithms for Clustering Data, *Prentice Hall*, 1988. 320 p.

[10] G. Karypis, E. Han, V. Kumar. Chameleon: A Hierarchical Clustering Algorithm Using Dynamic Modeling, *Computer* 32 (1999). p. 68−75.

[11] C. Ding, X. He. Cluster Aggregate Inequality and Multilevel Hierarchical Clustering, *Proc. 9th European Conf. Principles of Data Mining and Knowledge Discovery* (2005). p. 71−83.

[12] M. Laszlo, S. Mukherjee. A Gegridic Algorithm Using Hyper-Quadtrees for Low-Dimensional K-means Clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28 (4) (2006). p. 533−543.

[13] A. Vailaya, A.K. Jain, H.J. Zhang, “On image classification: city vs. landscape”, *Pattern Recognition*, vol. 31, p. 1921-1935, 1998.

[14] M.J. Swain,D.H. Ballard, “Color indexing”, *International journal of Computer Vision*, vol. 7, n. 1, p. 11-32, 1991.

[15] H. Nezamabadi-pour, E. Kabir, “Image retrieval using histograms of unicolor and bicolor blocas and direccional changes in intensity gradient”, *Pattern Recognition Letters*, vol. 25, n. 14, p. 1547-1557, 2004.

[16] F. Mokhtarian, S. Abbasi, “Shape similatity retrieval under affine transforms”, *Pattern Recognition*, vol. 35, p. 31-41, 2002.

[17] A.K. Jain, A. Vailaya, “Image retrieval using color and shape”, *Pattern Recognition*, vol. 29, n. 8, p. 1233-1244, 1996.

[18] B.S. Manjunath, W.Y. Ma, “Texture feature for browsing and retrieval of image data”, *IEEE PAMI*, vol. 8, n. 18, p. 837-842, 1996.

[19] J.R. Smith, C.S. Li, “Image classification and quering using composite region templates”, *Academic Press, Computer Vision and Understanding*, vol. 75, p. 165-174, 1999.

[20] J.Z. Wang, J. Li, G. Wiederhold, “SIMPLIcity: semantic sensitive integrated matching for picture libraries”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, n. 9, p. 947-963, 2001.

[21] R.A. Melnyk, R.B. Tushnytskyy, “Pattern Analysis by Clystering”, *Proc. of the 5th Intern. Conf. Neural Networks and Artificial Intelligence (ICNNAI)*(2008). p. 160–163.

Roman Melnyk *Professor of Software Engineering Department, Institute of Computer Sciences and Information Technologies, Lviv Polytechnic National University.*

Research interests: combinatory problems, methods of fuzzy clustering, coding and recognition, storage and data mining.

Ruslan Tushnytskyy *– a post-graduate student at the Software Engineering Department, Institute of Computer Sciences and Information Technologies, Lviv Polytechnic National University.*

Research interests: preliminary processing of images, coding and recognition in real time, clustering algorithms, data compression.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

ВИЗНАЧЕННЯ ЕЛЕМЕНТІВ ТЕКСТУ НА ЗОБРАЖЕННІ З ВИКОРИСТАННЯМ НЕЧІТКИХ КОГНІТИВНИХ МОДЕЛЕЙ

Олена П'ятикоп

Приазовський державний технічний університет
вул. Університетська, 7, Маріуполь 87500 Україна, E-mail: Pjatikopelena@rambler.ru

Резюме: *в даній публікації описуються результати моделювання знань з когнітивної психології про первинну обробку інформації в зоровій системі для аналізу зображення з метою визначення елементів тексту. Обробка зображення на рівні клітин первинної зорової кори формалізована за допомогою нечітких множин.*

Ключові слова: *зорова система, клітини первинної зорової кори, орієнтаційні колонки, зображення, рядки літери.*

ВСТУП

У галузі штучного інтелекту актуальним завданням залишається проектування систем автоматичної обробки графічної інформації. Одним з напрямків застосування таких систем є розпізнавання тексту. Вирішення цієї задачі актуально при перетворенні паперових схем, креслень та інших документів в їх електронний еквівалент. На сьогоднішній день вже є досвід розробки методів розпізнавання друкованих символів із застосуванням різного виду класифікаторів, нейронних мереж [1-2]. Але обсяг цифрової графічної інформації збільшується, змінюються умови розпізнавання, і від сучасних систем потрібно більш глибокий інтелектуальний аналіз. Тому для обробки зображення досліджується можливість застосування знань про будову зорової системи та процеси, що відбуваються в неї [3-5]. Сприйняття зорової та звукової інформації, а також процеси мислення і пам'яті описує когнітивна психологія [6-7].

Ці знання дають можливість моделювати сприйняття ока людини і застосовувати отримані моделі для обробки зображення.

У даній статті розглядають використання когнітивних моделей, зокрема моделі гангліозних клітин та моделі клітин первинної зорової кори для вирішення задачі визначення елементів тексту на зображенні.

1. ПОСТАНОВКА ЗАДАЧІ

Нехай є зображення з текстом в градаціях сірого 8 біт. Текст звичайно формується з рядків, а рядки, в свою чергу, складаються з слів. Тому спочатку необхідно визначити місце розташування передбачуваних рядків, далі розбити рядки на можливі слова, а кінцевим етапом виконати виділення букв. Розташування рядків на зображенні можливо не тільки горизонтальне, а під будь-яким кутом. Крім цього невідомий розмір букв тексту, тобто висота рядків. У зв'язку з цим, початковий аналіз зображення спрямований на визначення кута нахилу і висоти рядків тексту. Після визначення цих параметрів передбачуваний рядок розташовуємо горизонтально і виконуємо подальший аналіз.

Для дослідження зображення використовувати нечітку модель клітин первинної зорової кори.

Когнітивні знання

Якщо кинути побіжний погляд на зображення, не фіксуючи його на деталях, то відбувається сприйняття найбільш простих (укрупнених) паттернів. У розглянутій задачі це будуть рядки тексту, де рядки – це чергуються смуги одного напрямку і приблизно однієї ширини. Так відбувається, якщо зображення потрапляє в область периферичного зору, а також, якщо дивитися на зображення побіжно або здалеку. Після проецювання зображення на сітчатку в ній відбуваються такі процеси:

1) Першими образ на сітчатці обробляють фоторецептори палички і колбочки, які через проміжні біполярні клітини активізують певні рецептивні поля гангліозних клітин. Розрізняють

гангліозні клітини двох типів: "ON" клітини активізуються, коли світло потрапляє на деякий чутливий центр, і не реагують, коли світло падає на область навколо цього центру; клітини "OFF" типу реагують навпаки. Тому ці клітини ще називають релейними [6-7]. Безліч активних гангліозних клітин утворює першу "проекцію" зображення.

2) Далі, більшість аксонів гангліозних клітин утворюють сінаптичні зв'язки з клітинами латерального колінчастого тіла (ЛКТ), рецептивні поля яких дуже схожі на рецептивні поля гангліозних клітин сітчатки: мають центральні "зони включення" і периферійні "зони виключення" чи навпаки. Крім цього клітини ЛКТ відрізняються розмірами і діляться на дві групи: парвоцеллюлярні і магноцеллюларні клітини.

3) В ЛКТ перша "проекція" зображення стає пульсуючою [8]. Відразу ж після саккади[1] в ЛКТ формується ретинотопічна карта, яка передається в первинну зорову кору. Діаметр рецептивних полів зменшується, і на основі нього формуються нові ретинотопічні карти. До чергового стрибка зорова кора переробляє дані, отримані з ЛКТ.

4) Для збудження клітин зорової кори потрібно більш тонкий механізм, що відображається на анатомічних особливостях самих клітин і на сигналах, які необхідні для їх збудження. Існує декілька типів цих клітин: *прості* клітини реагують тільки на лінійні сегменти, орієнтовані певним чином; *складні* клітини вимагають руху в певному напрямку; *гіперскладні* клітини вимагають, щоб стимули, що потрапляють у середину їх рецептивних полів, були певної довжини. Окрім, цього відомо [7,9], що клітини, які реагують на одну орієнтацію, формують, собою колонку зорової кори (рис. 1). При цьому рецептивні поля прилеглих клітин мають інші орієнтаційні уподобання, які змінюються поступово. Таким чином, серед безлічі орієнтаційних колонок, на рядки тексту найбільш активно проявить себе одна колонка (або група суміжних) певної орієнтації. В цьому випадку множина клітин цієї колонки буде відповідати, в залежності від довжини, або рядку тексту, або вже конкретному слову. Тоді кут нахилу рядків буде відповідати куту обраної колонки, а висота рядка – розмірам ядра клітини.

Формальний опис моделі гангліозних клітин та можливість її застосування докладно розглянуто в [9]. В даній публікації вже розглядається застосування моделі клітин первинної зорової кори.

2. МОДЕЛЬ КЛІТИН ПЕРВИННОЇ ЗОРОВОЇ КОРИ

Оскільки для активізації кліток цього рівня необхідна активність групи суміжних гангліозних кліток [8], то введемо поняття детектора. Сукупність суміжних кліток показана на рис. 1.

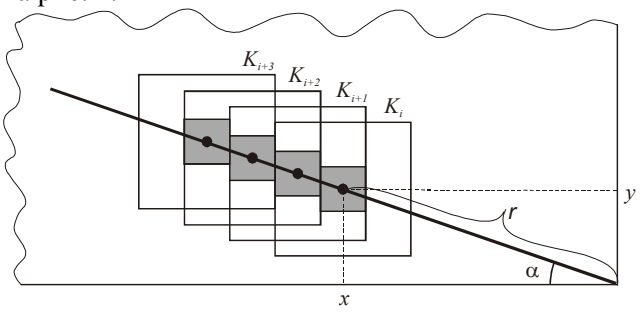


Рис. 1 – Сукупність суміжних клітин

Тоді модель детектора це вираз (1):

$$D^{\alpha} =< K^{\alpha,r}, \alpha, \omega, \ell^{*}, \delta(D^{\alpha}) >, \quad (1)$$

де $K^{\alpha,r} =< K^{\alpha,r+2Rk}, K^{\alpha,r+4Rk} \dots K^{\alpha,r+\ell^{*}\cdot 2Rk} >$;

α – кут орієнтації від 0^0 до 175^0;

ω – ширина детектора, що визначається $\omega = Rk$;

ℓ^{*} – довжина детектора, що визначається кількістю клітин суміжних, $K^{a,r}$, але не менш 4;

$\delta(D^{\alpha})$ – функція впевненості наявності детектора, обчислюється за формулою (2):

$$\delta(D^{\alpha}) = \begin{cases} 0, & \text{якщо } \prod_{i \in 1 \dots \ell^{*}} \mu(K_i) = 0 \\ \frac{1}{\ell^{*}} \cdot \sum_{i \in 1 \dots \ell^{*}} \mu(K_i), & \text{якщо } \prod_{i \in 1 \dots \ell^{*}} \mu(K_i) > 0 \end{cases} \quad (2)$$

де $\mu(K_i)$ де обчислюється за формулами [9].

На одному проведеному промені може бути кілька детекторів. Нехай кількість цих детекторів буде равно n_{α}. Тоді множина всіх детекторів буде $D^{\alpha} = \{D_j^{\alpha}\}$, $j \in 1 \dots n_{\alpha}$, що визначаються виразом (1).

На основі поняття детектора опишемо модель різних видів клітин зорової кори.

Так, проста клітина S^{α}, що реагує тільки на певну орієнтацію, описується виразом (3):

$$S^{a} =< K^{\alpha,r}, \alpha, \omega, \ell, \delta(S^{\alpha}) >, \quad (3)$$

[1] рухи очей, які переводять точку фіксації з однієї ділянки зображення на інший, та використовуються переважно для обстеження та вивчення поля зору

де $\ell \geq \ell^*$, а $\delta(S^\alpha) = \min\limits_{j \in 1 \ldots n_\alpha} \{\delta(D_j^\alpha)\}$.

Складні клітини розглянуті не будуть, оскільки вони реагують на рух, а в даній роботі розглядається обробка статичного зображення.

Гіперскладні клітини зорової кори, також як і прості, чутливі до напрямку сигналу, але суттєвої їх особливістю є реакція на певну довжину ℓ. Тому гіперскладна клітина $G^{\alpha,\ell}$ може бути описана виразом (4):

$$G^{\alpha,\ell} =< K^{\alpha,r}, \alpha, \omega, \ell, \delta(G^{\alpha,\ell}) >, \quad (4)$$

де $\delta(G^{\alpha,\ell}) = \delta(D^a)$.

В одному напрямку α гіперскладних клітин однієї довжини може бути кількість m_α. Тоді множина всіх гіперскладних клітин буде описано $G^{\alpha,\ell} = \{G_j^{\alpha,\ell}\},\ j \in 1 \ldots m_a$. Але також може бути довжини $\ell_1, \ell_2, \ldots \ell_z$. Тоді множина гіперскладних клітин, реагуючих на певні довжини, висловлюється так

$$G^{\alpha,\ell_i} = \{G_j^{\alpha,\ell_i}\},\ i \in 1 \ldots z,\ j \in 1 \ldots m_a$$

У підсумку можна побудувати модель зображення у вигляді безлічі орієнтаціонних колонок, де кожна колонка описується виразом (5):

$$C^\alpha = \left\{S^\alpha, \{G_j^{\alpha,\ell_i}\}_{\ell_i}\right\}, \alpha = 0,5,10 \ldots 175 \quad (5)$$

Тоді модель "проекції" зображення, тобто ретинотопічної карти описується виразом (6):

$$M^\omega =< \{C^\alpha\}_\omega, \omega, \lambda(M^\omega) >, \quad (6)$$

де α – кут орієнтації від 0^0 до 175^0;

ω – ширина детектора, що визначається ω= *Rk*;

$\lambda(M^\omega)$ – функція впевненості наявності найбільшої кількості гіперсложних клітин найбільшої довжини, обчислюється за формулою (7):

$$\lambda(M^\omega) = [\frac{1}{2}(\lambda_1 + \lambda_2)]_{\alpha^*}, \quad (7)$$

де λ_1 – величина відображення наскільки всі активні релейні клітини сприяли активізації гіперсложних клітин;

λ_2 – середнє значення впевненості всіх гіперсложних клітин шириною ω;

α^* – кут орієнтації від 0 до 175, обраний на множині оріентаційних колонок C^α.

Тоді λ_1 визначається виразом (8):

$$\lambda_1 = \frac{n^{K^\omega}}{n^{G^{\alpha,\ell,\omega}}}, \quad (8)$$

де n^{K^ω} - кількість активних гангліозних клітин K^ω з розміром ядра $Rk = \omega$;

$n^{G^{\alpha,\ell,\omega}}$ - кількість активних гангліозних клітин з розміром ядра $Rk = \omega$, які активізували гіперскладні клітини $\{G^{\alpha,\ell_i}\}_{\ell_i}$.

У свою чергу λ_2 визначається як (9):

$$\lambda_2 = \frac{1}{m_\alpha} \sum_j \sum_i \delta(G_i^{\alpha,\ell_j,\omega}), \quad (9)$$

де m_α – загальна кількість гіперскладних клітин одного напрямку, а кут α^* визначається з виразу $l_{\alpha^*} = \max\limits_\omega \{\max\limits_\alpha \{l_\alpha\}\}$, де визначається для α від 0 до 175.

Після визначення для кожної ретинотопічної карти M^{ω_i} її функції впевненості потрібно вибрати карту з найбільшою упевненістю. Тоді клітини цієї карти будуть відповідати або смугам, або словам висотою ω під кутом α на зображенні.

3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Розглянемо приклад застосування наведених моделей за допомогою зображення розміром 860×227, яке наводиться на рис. 2. На рис. 3 показана ретинотопічна карта, отримана в результаті автоматичного визначення оптимального кута нахилу $\alpha=5^0$ і ширини ω=12 пікселів передбачуваних рядків тексту.

Детальний процес вибору оптимальних параметрів приводиться в роботі [10].

Оскільки моделювання клітин прив'язано до піксельного полю, то знаючи координати розміщення клітин, можна локалізувати і елементи, які вони утворюють, що показано на рис. 4.

Завдяки класичним формулам комп'ютерної графіки (10) та визначеному куту можно виконати поворот зображення.

$$x' = x \cdot \cos(\alpha) - y \cdot \sin(\alpha)$$
$$y' = x \cdot \sin(\alpha) + y \cdot \cos(\alpha) \quad (10)$$

де (x, y) – початкові, а (x`,y`) – кінцеві координати.

Результат зображення після повороту наведено на рис.5.

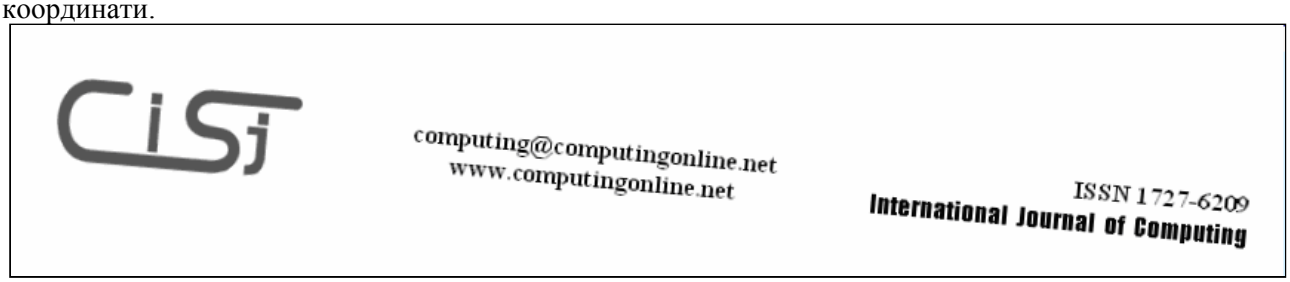


Рис. 2 – Початкове зображення

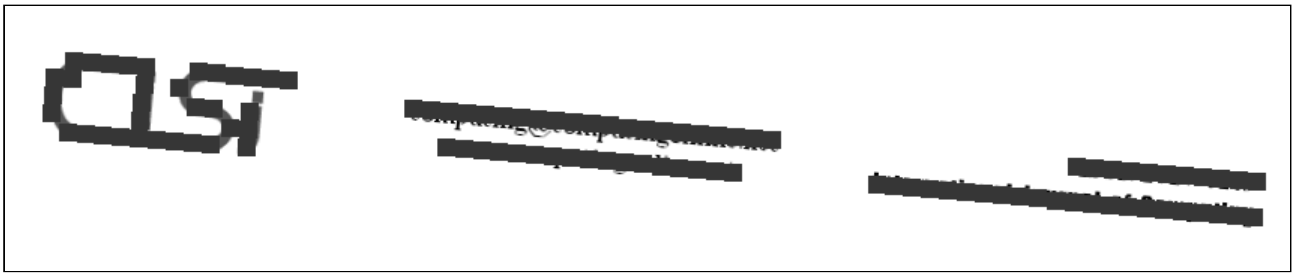


Рис. 3 – Ретинотопічна карта, отримана в результаті автоматичного визначення оптимального кута нахилу $\alpha=5^0$ і ширини $\omega=12$ пікселів можливих рядків тексту.

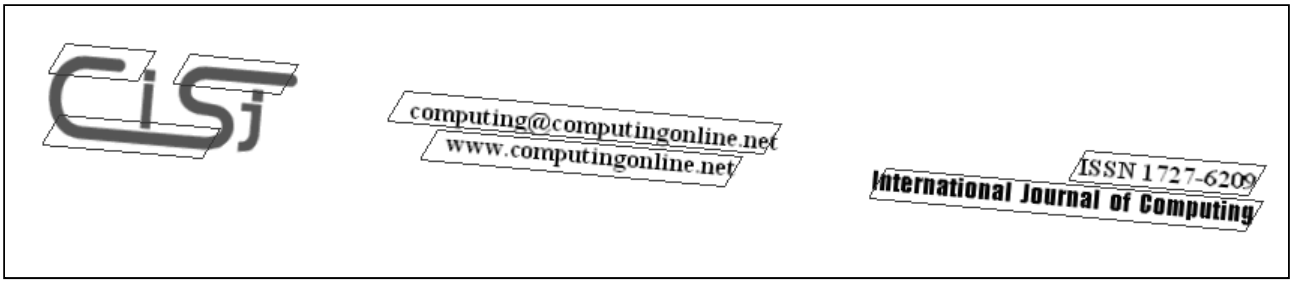


Рис. 4 – Результат локалізації елементів можливих строк тексту.

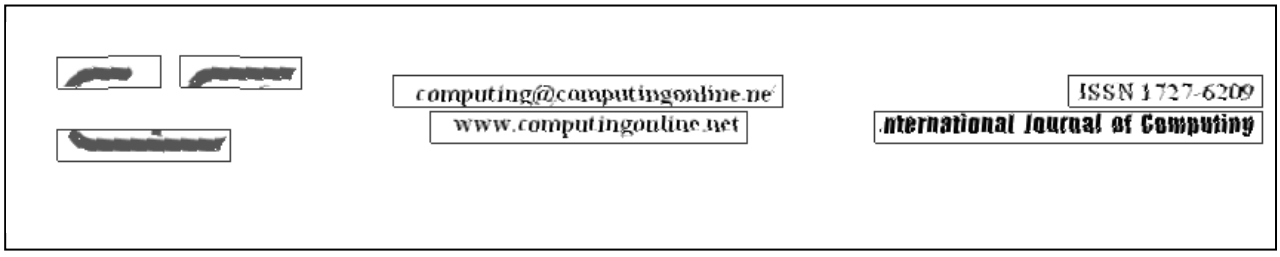
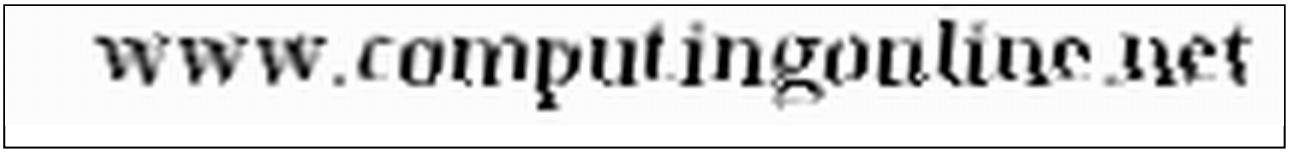


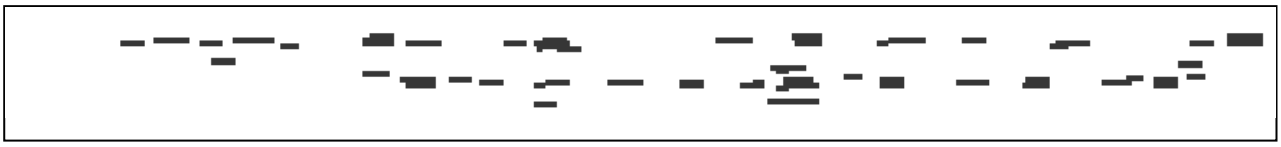
Рис. 5 – Зображення після повороту.

Тепер, коли основні крупні елементи зображення локалізовані, можна виконати більш детальний аналіз, тобто наступна саккада стосується кожного окремого елементу. Можна спробувати розбити кожний із елементів на літери, а можна попередньо проаналізувати, чи має цей елемент характерні для слів особливості. Під характерними особливостями слів мається на увазі наявність послідовності вертикальних паттернів, та низки горизонтальних [11]. Для цього треба знову представити зображення за допомогою моделі клітин первинної зорової кори тільки зі зменшеним рецептивним полем.

На рис. 6 наведено приклад наявності таких характерних особливостей, а на рис. 7 – їх відсутність. На рис. 8 показаний наступний крок –розбивка на літери, який виконується теж на основі представлення зображення за допомогою клітин первинної зорової кори. Як видно з рис. 8 в трьох крайніх лівих елементах розбивка на літери не виконувалась. Подальша обробка їх також не стосується.



а) початкове зображення



б) горизонтальні елементи

в) вертикальні елементи

Рис. 6 – Наявність характерних для слів особливостей

а) початкове зображення

б) горизонтальні елементи

в) вертикальні елементи

Рис. 7 – Відсутність характерних для слів особливостей

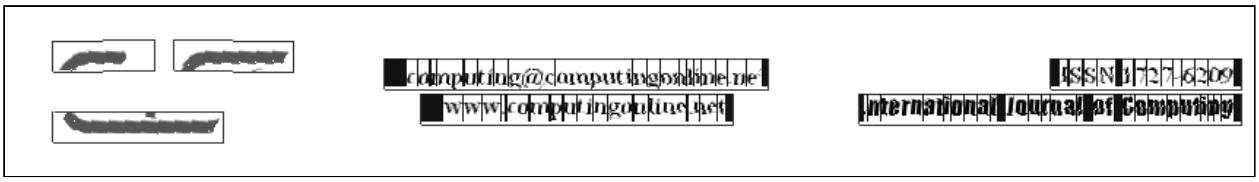


Рис. 8 – Розбивка на літери

Далі для кожного з отриманих блоків необхідно виконати процедуру розширення області для надрядкових або підрядкових елементів, а також процедуру обрізки зайвого фону. Після цього, блок-обводка буде чітко відповідати ширині і висоті символу (символів). На сьогоднішній день досить досліджені ці співвідношення між висотою і шириною для різних символів та шрифтів. На підставі цього приймемо:

1) якщо це співвідношення більше 0,85 то, швидше за все, це один символ;
2) якщо співвідношення висоти до ширини більше 0,65, але менше 0,85, то припустимо, що це може бути один із символів ш, щ, м, ж, ю, або група символів;
3) якщо співвідношення висоти до ширини менше або дорівню 0,65, то це група декількох символів.

З аналізу символів першої групи можна встановити середнє значення ширини символів, і використовуючи ці дані умовно розбити блоки третьої групи. Кінцевий результат локалізації літер показаний на рис. 9.

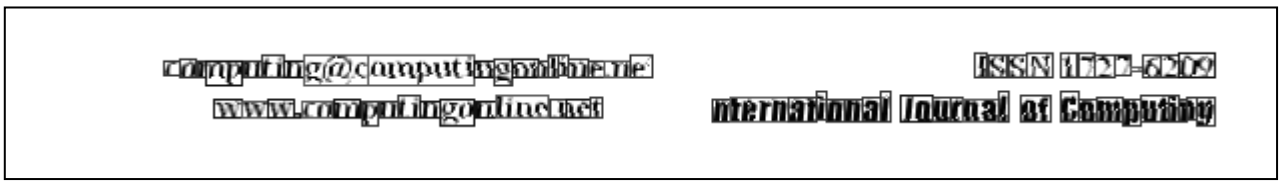


Рис. 9 – Локалізація літер

5. ВИСНОВКИ

Проведені дослідження підтвердили можливість використання знань про будову зорової системи, а саме про первинну обробку інформації на рівні клітин зорової кори, для обробки зображення з метою виявлення елементів тексту. Використання розробленої моделі клітин первинної зорової кори дозволяє представити зображення локалізованого символу у вигляді множини елементів, подальше групування яких дозволить виконати ідентифікацію.

6. СПИСОК ЛІТЕРАТУРИ

[1] В. Л. Арлазаров. *Адаптивное распознавание символов* [Электронный ресурс]/ В.Л. Арлазаров, В.В. Троянкер, Н.В. Котович //Режим доступа к статье: http://www.cuneiform.ru/tech/adaptive.html

[2] А. С. Васюра. *Моделирование нейросети для решения задачи идентификации символов* [Электронный ресурс]/ А.С. Васюра, Т. Б. Мартынюк, Л. М. Куперштейн // Наукові праці ВНТУ – 2007 – № 1 – Режим доступа к статье: www.nbuv.gov.ua/e-journals/VNTU/2007-1/ru/07vasosi_ru.pdf

[3] Е. И. Ярмошевич. *Функциональная спектральная пространственно-временная модель формирования изображений объектов зрительной системой человека* [Электронный ресурс]/ Е. И. Ярмошевич, Е. Е. Михайлова, М. А. Пономаренко // Вестник ВГУ, серия: системный анализ и информационные технологии. – 2008. – №1 – стр. 74-78. – Режим доступа к статье: – www.vestnik.vsu.ru/pdf/analiz/2008/01/yarmoshevich.pdf

[4] J.A. Bednar. *Scaling Self-Organizing Maps To Model Large Cortical Networ*ks [Электронный ресурс], Neuroinformatics, 2:275-302, 2004. Режим доступа: http://nn.cs.utexas.edu/keyword?bednar:neuroinformatics04 – название экрана

[5] С.А. Гладилин. *Нейронная сеть, воспроизводящая выходной сигнал ганглиозной клетки* [Электронный ресурс]/ С.А. Гладилин, Д.Г. Лебедев // Информационные процессы – 2005.– том 5, №3.– стр. 258-264. – Режим доступа к статье: http://www.jip.ru/2005/258-264.pdf

[6] Р.Л. Солсо. *Когнитивная психология*, СПб.: "Питер", 2002. – 592с.

[7] Х.Р. Шиффман. *Ощущение и восприятие*, СПб.: "Питер", 2003. – 928с., ил.

[8] В.Е. Демидов. *Как мы видим то, что видим*, М.: Знание, 1987– 240 с.

[9] О. П`ятикоп. Застосування когнітивних моделей сприйняття зображення в задачах розпізнавання тексту, *Proceedings of the III Conference on Computer Science and information Tecnologies (CSIT`2008), Lviv, Ukrain, September 25-27,2008, pp.168-171*

[10] Е.Е. Пятикоп. Некоторые результаты компьютерных экспериментов локализации строк текста на основе когнитивных моделей восприятия изображения, *Вісник донецького університету*, *Сер. А: Природничі науки,* Донецьк, 2008., № 2, ч.2, стр. 527-532.

[11] Zhang L, Lu Y, and Tan C L. Italic Font Recognition Using Stroke Pattern Analysis on Wavelet Decomposed Word Images, *The 17th International Conference on Pattern Recognition*, ICPR2004, Cambridge, United Kingdom, 23-26 August 2004, vol.4, pp. 835-838 // Режим доступа к статье www.comp.nus.edu.sg/~tancl/Papers/ICPR04/Italic_DIAL2004.pdf

Олена П'ятикоп, *в 2001 році отримала диплом з спеціальності "Інформатика" Приазовського державного технічного університету м. Маріуполь, працює асистентом на кафедрі інформатики цього навчального закладу, навчається в аспірантурі.*

Наукові інтереси: методи штучного інтелекту, нечітка логіка, комп'ютерна графіка.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

DETERMINING THE ELEMENTS OF THE TEXT ON IMAGE USING OF FUZZY COGNITIVE MODELS

Elena Pyatikop

Priazov State Technical University
7, Universitetska Street, Mariupol 87500 Ukraine, E-mail: Pjatikopelena@rambler.ru

Abstract: *This publication describes the modeling results of knowledge of cognitive psychology about the primary processing in the visual system to analyze the images in purpose to determine the elements of the text. Processing images at the level of primary visual cortex formalized using fuzzy sets.*

Keywords: *visual system, cells of the primary visual cortex, orientation columns, images, lines of letters.*

INTRODUCTION

In the field of artificial intelligence the challenge task is recognize text. Today, there is experience in developing methods of recognizing printed characters with different kinds of classifiers, neural networks [1-2]. But the volume of digital image information is increased the conditions of recognition are change and modern systems require a more in-depth intellectual analysis. Therefore for image processing the possibility of applying knowledge about the structure of the visual system and the processes occurring in it is explored [3-5]. The cognitive psychology describes this knowledge [6-7]. This article discusses the use of cognitive models, in particular the model ganglion cells and cell models of primary visual cortex to solve the identity of the elements of the text in the image.

1. THE PROBLEM

Suppose there is an image with text in 8 bit grayscale. The location of lines in the image may be at any angle. In addition, the letter size of the text is unknown. In this regard, an initial analysis of the image is intended to define the angle of slope and height of text lines. After that the line must be divided into possible words, and then make the selection of letters.

To investigate the images you need to use fuzzy model of the image cells of primary visual cortex.

If you glance at the image not fixing it on details, it will produce the perception of larger patterns. In this problem it will be a text lines. That is, if the image falls in the area of peripheral vision. After projecting the image on retina it is occurred the following processes:

1) for the first the image is processed by rods and cones. These rods and cones activate some receptive fields of ganglion cells through intermediate bipolar cells [6-7]. A lot of active ganglion cells form the first “projection” images.

2) next, the majority of ganglion cell axons form synaptic couplings with cells lateral geniculate nucleus (LGN). In the result it is an exact repetition of the retina in LGN, i.e. its retina “map”. The cells of this level by the structure of buildings are like the ganglion cells, but also differs by sensitivity and sizes.

3) In turn, LGN axons transmit signals to the primary (striate) visual cortex. The cells of visual cortex respond to specific orientation and the length of the stimulus. Cells that are responded to the some location of the stimulus are the orientation column. Thus it is formulated about 18-20 columns, with a gradual change in orientation from 0^0 to 180^0.

In that way, the most active column will reflect the lines of text. A lot of cells of this column are related to the whole line or the separate words of this one. The angle of this active column corresponds with the angle of slope of the text lines. And the size of the cell nucleus corresponds with the height of the line.

Formal description of the model ganglion cells and the possibility of its usage are discussed in detail in [9]. But in this publication it is considered the usage of models of primary visual cortex cells.

2. CELL MODEL OF PRIMARY VISUAL CORTEX

As for the activation of cells at this level must be

closely related group of ganglion cells [8], for that we introduce the concept of the detector. The combination of adjacent cells is shown in Figure 1. Then the model of the detector, is the expression (1), where $K^{\alpha,r} =< K^{\alpha,r+2Rk}, K^{\alpha,r+4Rk} \dots K^{\alpha,r+\ell^*\cdot 2Rk} >$; α – angle of orientation 0^0 to 175^0; ω – width of the detector, which is defined $\omega = Rk$; ℓ^* – length detector, which is determined by the number of adjacent cells $K^{a,r}$, but not less 4; $\delta(D^\alpha)$ – the confidence function of the detector presence is calculated by the formula (2), where $\mu(K_i)$ is calculated by the formula [9].

Set of detectors will be $D^\alpha = \{D_j^\alpha\},\ j \in 1 \dots n_\alpha$ Based on the concept of the detector we will describe a model of different types of cells of the visual cortex.

Thus, a simple cell S^α, which responds only to a certain orientation, described by expression (3), where $\ell \geq \ell^*$, but $\delta(S^\alpha) = \min\limits_{j \in 1 \dots n_\alpha} \{\delta(D_j^\alpha)\}$.

The most complicated visual cortex cells, as well as simple, are sensitive to the direction of the signal. But their basic characteristic is a reaction to a certain length ℓ. Then the most complicated visual cortex cell $G^{\alpha,\ell}$ can be described by the expression (4), where $\delta(G^{\alpha,\ell}) = \delta(D^a)$.

In one direction α the most complicated visual cortex cells by the same length can be quantity m_α. Then the set of all the most complicated cells will be described $G^{\alpha,\ell} = \{G_j^{\alpha,\ell}\},\ j \in 1 \dots m_a$. But these may be another length $\ell_1, \ell_2, \dots \ell_z$. Then the set the most complicated cells reacting to specific lengths, expressed as

$$G^{\alpha,\ell_i} = \{G_j^{\alpha,\ell_i}\},\ i \in 1 \dots z,\ j \in 1 \dots m_a$$

As a result, you can build a model of the image as a set of orientation columns, where each column is described by expression (5).

Then the model of "projection" image, i.e. retina map is described by the expression (6), where α – angle of orientation 0^0 to 175^0; ω – width of the detector, which is defined $\omega = Rk$; $\lambda(M^\omega)$ – the confidence function of the existence of the greatest number of the most complicated cells of the maximum length is calculated by the formula (7), where λ_1 – is the value of all active relay cells have enhanced to the most complicated cells is defined by (8); λ_2 – average value of the confidence of all the most complicated cell with width ω, is defined (9); α^* – angle of orientation 0 to 175, selected on the set of orientation columns C^α, determined by the expression $l_{\alpha^*} = \max\limits_{\omega}\{\max\limits_{\alpha}\{l_\alpha\}\}$.

You need to calculate the confidence function for each retina map retina map M^{ω_i} to select the retina map which has the maximum confidence function.

Then the cells of this map will correspond the whole lines or the separate words of the lines with angle α and height ω.

3. RESULTS

Consider an example application of the previous models at an image size 860 × 227, which is given in Fig. 2. Fig. 3 shows retina map obtained from the automatic determination of optimal angle of inclination. Detailed process of selecting the optimal parameters is given in [10]. Fig. 4 shows the localization of the basic elements of the text lines. Using the classical transformation of computer graphics and a calculated angle, you can rotate the image. And the result is shown in Fig. 5.

When the basic larger elements of images are located, you can make more detailed analysis. The following movement of the eye relates to each individual element and has a projection on the retina.

You can try to split each of the elements in the letters, but you can pre-examine whether this element has particular word characteristics. Under the features of the words it is implied the existence of a sequence of patterns of vertical and horizontal series [11]. To do this you need to re-submit images using models of cells of primary visual cortex only with a reduced receptive field. On Fig. 6 there is an example of the availability characteristics to the letters, as in Fig. 7 - their absence. Fig. 8 illustrates the next step - a breakdown in the letter, which is performed also on the basis of the images using the cells of primary visual cortex.

To each of the blocks need to perform the procedure of expanding the field for superscript or subscript elements of the letters. As well as you need make the procedure for cutting off the extra background. After that, the frame will closely match the width and height of character (s). The final result of localization of letters is shown on Fig.9.

4. CONCLUSION

The results of investigations have confirmed the possibility of applying knowledge about the structure of the visual system for processing image in aim to determine the elements of the text.

The usage of developed models of cells of primary visual cortex permits to show the image of located symbol as the set elements. Each element is characterized by its own length and orientation. They are the input data for the further identification symbol.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

RESEARCHES OF THE COMBINATORIAL MODELS FOR INNOVATIVE INFORMATION TECHNOLOGIES

Volodymyr Riznyk [1) 2)]

[1)] Lviv Polytechnic National University, 12 S.Bandera Str., 79013, Lviv, Ukraine, rvv@polynet.lviv.ua, http://iknit.lp.edu.ua/riznyk
[2)] University of Technology and Life Sciences Bydgoszcz, 7 Kaliskiego Ale, 85-796, Poland, wriz@utp.edu.pl

Abstract: *The paper presents a new mathematical principle of innovative techniques for improving the quality indices of engineering devices and systems with non-uniform structure (e.g. coding systems) with respect to transmission speed, positioning precision, resolving ability, and functionality, using novel design based on remarkable properties and structural perfection of one- and multidimensional models of the systems, namely the concept of Ideal Ring Bundles (IRB)s prospected from basic laws of the world-wide harmony. Research into the underlying mathematical principle provides an ability to reproduce the maximum number of combinatorial varieties in the systems with a limited number of elements and bonds. This approach make it possible to configure systems with optimal placement of structural elements in spatially or temporally distributed systems, using the appropriate mathematical apparatus of contemporary combinatorial theory.*

Keywords: *Combinatorial model, Information technology, Numerical model, Coding system, Transmission speed, Resolving ability, Multidimensional model, Vector data coded design, Ideal Ring Bundle, World-wide harmony.*

1. INTRODUCTION

Problem of structural optimization in systems engineering relating to finding the best placement both elements and bonds of the system as well as improving its quality indices, is known, to be connected not only with number theory and combinatorial analysis but also with research of the fundamental laws of real world, such as symmetry laws [1], [2] and aurea section [2]. Research into underlying mathematical area involves investigation of novel techniques based on Combinatorial Mathematics, and Combinatorial Sequencing Theory, namely the concept of Ideal Ring Bundles [3], which can be used for finding optimal solutions for some problems in systems engineering and information technologies.

2. NUMERICAL MODEL OF HIGH-PERFORMANCE SYSTEMS

The ordered chain approach to the study of elements and events is known to be of widespread applicability when applied to the problem of finding the optimum ordered (non-redundant) arrangement of structural elements in a distributed technological or computing system. Let us regard an n-stage sequence of distinct positive integers $K_n = \{k_1, k_2, \ldots k_n\}$ as being cyclic, so that k_n is followed by k_1. We call this a ring sequence (Fig.1). A sum of consecutive terms in the ring sequence can have any of the n terms as its starting point, and can be of any number of terms from 1 to n-1. In addition, there is the sum of all n terms, which is the same independent of the starting point. Hence, the maximum number of distinct sums S_n of consecutive terms of the ring sequence is given by:

$$S_n = n(n\text{-}1) + 1 \qquad (1)$$

Here is graphical model of an n-stage ring sequence $K_n = \{k_1, k_2, \ldots k_n\}$(Fig.1).

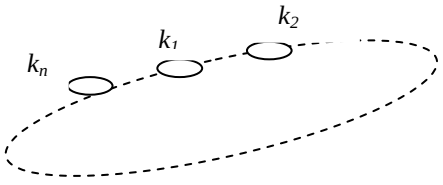


Fig.1 – Graphical model of an n-stage ring sequence $K_n = \{k_1, k_2, \ldots k_n\}$

If we require all terms in each sum to be consecutive elements of the sequence, for which the set of all S_n circular sums consists of the numbers from 1 to $S_n = n(n\text{-}1)+1$ that is each number occurs exactly once is called *Ideal Ring Bundle* (IRB) with

parameters n, S_n. Here is an example of an IRB with parameters n=4, S_n =13, namely {1, 3, 2, 7} (Fig.2).

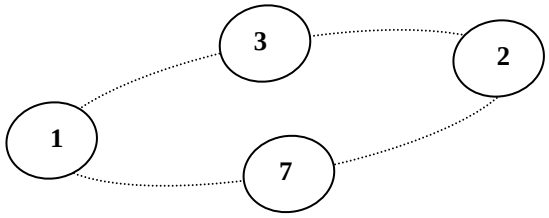


Fig. 2 – The Ideal Ring Bundle with n=4, S_n =13

To see this, we observe complete set of circular sums calculated from the IRB {1,3,2,7} (Table 1).

Table 1. Circular sums of the IRB {1,3,2,7}

1=1	2=2	3=3	4=1+3
5=3+2	6=1+3+2	7=7	8=7+1
9=2+7	10=2+7+1	11=7+1+3	12=3+2+7
13=1+3+2+7			

Note that if we allow summing over more than one complete revolution around the ring, we can obtain all positive integers as such sums. Thus:

14 = 1+3 +2 +7 +1, 15 = 2 +7 +1+3 +2,
16 = 3+2 +7 +1 +3, etc.

One-dimensional (1-D) Ideal Ring Bundle is an ordered numerical construction with n distinct integers, which form perfect partitions of finite interval [1,s] of integers. The sums of connected sub-sequences of 1-D IRB enumerate the set integers [1,s] exactly R-times [3].

The favorable qualities of numerical models provide many opportunities to apply them to advanced information technology and numerous branches of science.

3. IRB CODING SYSTEMS

Concept of the Ideal Ring Bundles can be used for finding optimal solutions for some problems of information computational processes. For example, Ideal Ring Bundles make it possible to configure optimum coding system, based on so-called IRB Monolithic Binary Code. This code forms binary code combinations, which all symbols "1" as well as symbols "0" are arranged together [4]. The IRBs provide an ability to reproduce the maximum number of the "monolithic" code combinations (all symbols "1" as well as symbols "0" are arranged together) in the coding system with a limited number of code word digits. First of all we can see computational processes, based on the IRB Monolithic Binary Code.

Here is an example of one-dimensional Monolithic Binary Code based on the IRB {1,3,2,7} with parameters n=4, S=13:

Table 2. One-dimensional Monolithic Binary Code based on the IRB {7, 2, 3, 1}

Number	Code combination
0	0000
1	0001
2	0101
3	0010
4	0011
5	0110
6	0111
7	1000
8	1001
9	1100
10	1101
11	1011
12	1110
13	1111

Table 2 contains the complete set of binary code combinations for coding the numbers 0, 1,…13.

Underlying combinatorial constructions can be represented as mathematical model of optimum coding system, based on the IRBs conception.

4. GENERALIZED IDEAL RING BUNDLES

Next, we consider a more general type of IRB, where the S_n circular sums of consecutive terms give us each integer value from 1 to N, for some integer N, exactly R times, as well as the value $N - 1$ (the sum of all n terms) exactly once. Here we see that:

$$N = n(n\text{-}1)/R \qquad (2)$$

Here is a graph of the IRB with n = 4, R = 2 and N = 6, where k_1=1, k_2,=1, k_3=2, k_4=3} (Fig.3).

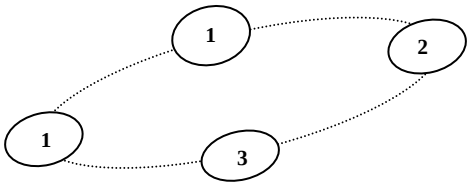


Fig. 3 – Graph of the IRB with n =4, N =6, R =2

Sums of consecutive terms for the IRB {1, 1, 2, 3} are:

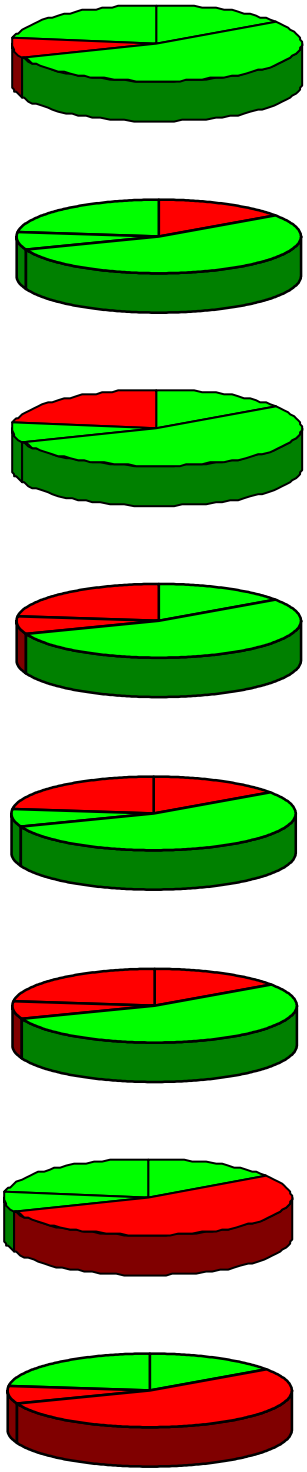
1:	1,	and	1,
2:	1 + 1,	and	2,
3:	1 + 2,	and	3,
4:	1 + 1 + 2	and	3 + 1,
5:	2 + 3	and	3 + 1 + 1,
6:	1 + 2 + 3	and	2 + 3 + 1

Using equation (2) easy to calculate sums of all n terms for double (R=2) IRB, namely: 2, 4, 7, 11, etc. We say this numerical set is generative double-IRBs

row.

5. SEGMENTATION OF IDEAL RING BUNDLES

Let us regard of the IRB{1,3,2,7} depicted above (Fig.2) and the IRB{1,1,2,3} (Fig.3) as graphic vision of an image segmentation of the IRBs (Fig.4).



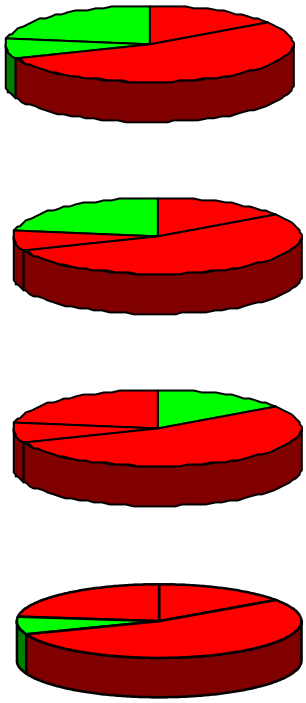
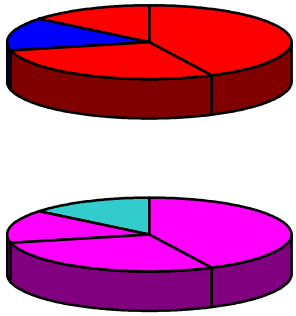


Fig. 4 – Graphic segmentation of the IRB {1,3,2,7}

Regarding picture (Fig.4), easy to see exhaustive set of harmonious two-body relationships (here is red and green) from 1:12 to 1:12 obtained exactly once (R = 1) each of them using the IRB {1, 3, 2, 7} image segmentation.

So, 13 is a lucky number, because it allows on the perfect partition 13=1+3+2+7 and provides an ability to reproduce the maximum number of harmonious two-body relationships. The fact makes it possible to develop of new approach in research of the discrete structures from epistemological point of view or “nature and nurture” [2].

Now, let us regard a graphic segmentation of the Ideal Ring Bundle {1, 1, 2, 3}. Here is exhaustive set of harmonious two-body relationships from 1:6 to 6:1 obtained exactly twice (R = 2) each of them over the IRB {1, 1, 2, 3} (Fig. 5).



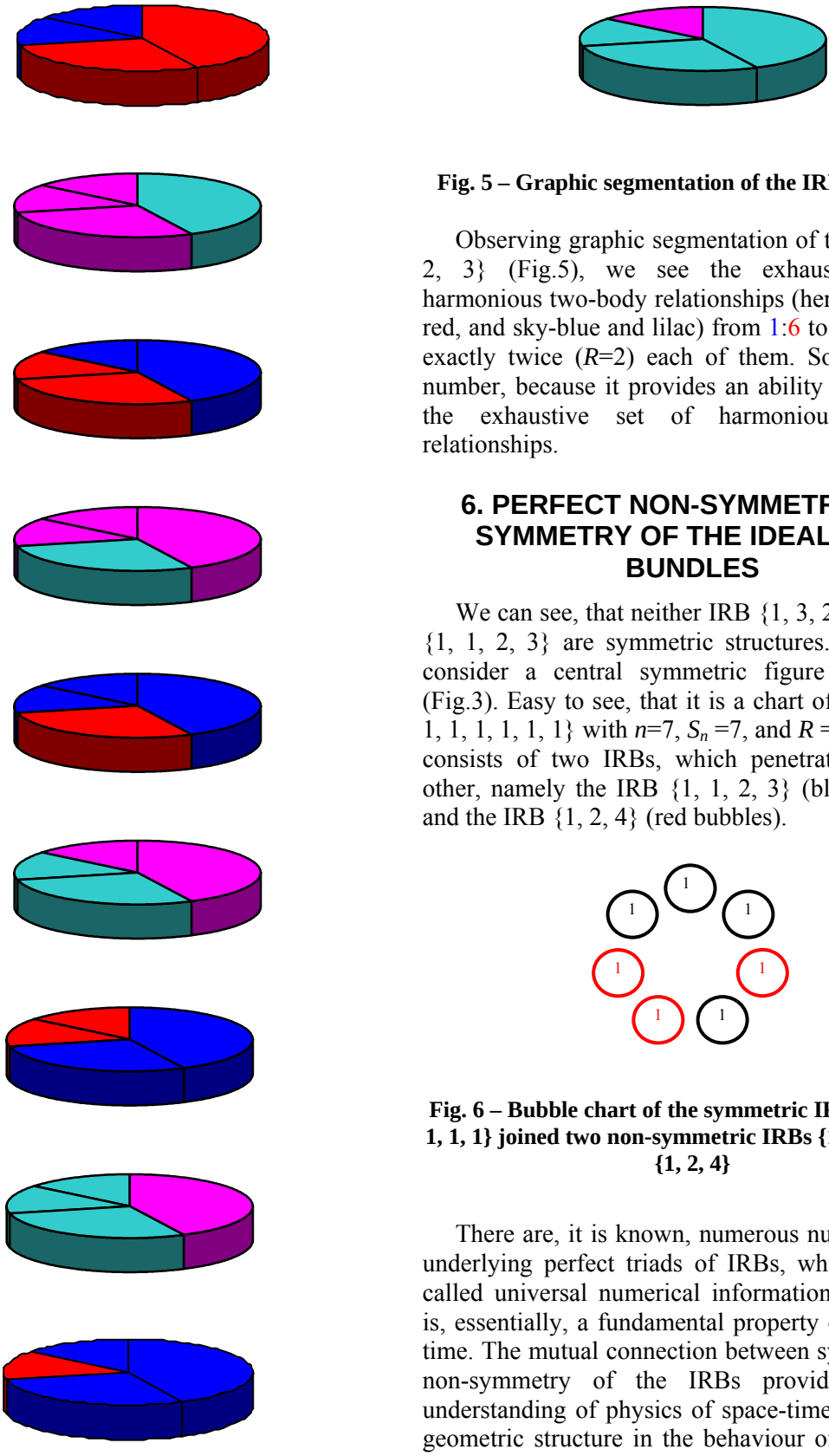


Fig. 5 – Graphic segmentation of the IRB {1, 1, 2, 3}

Observing graphic segmentation of the IRB{1, 1, 2, 3} (Fig.5), we see the exhaustive set of harmonious two-body relationships (here is blue and red, and sky-blue and lilac) from 1:6 to 1:6 obtained exactly twice (R=2) each of them. So, 7 is lucky number, because it provides an ability to reproduce the exhaustive set of harmonious two-body relationships.

6. PERFECT NON-SYMMETRY AND SYMMETRY OF THE IDEAL RING BUNDLES

We can see, that neither IRB {1, 3, 2, 7} nor IRB {1, 1, 2, 3} are symmetric structures. Now let us consider a central symmetric figure of order 7 (Fig.3). Easy to see, that it is a chart of the IRB {1, 1, 1, 1, 1, 1, 1} with n=7, S_n =7, and R =7. This chart consists of two IRBs, which penetrate with each other, namely the IRB {1, 1, 2, 3} (black bubbles) and the IRB {1, 2, 4} (red bubbles).

Fig. 6 – Bubble chart of the symmetric IRB {1, 1, 1, 1, 1, 1, 1} joined two non-symmetric IRBs {1, 1, 2, 3} and {1, 2, 4}

There are, it is known, numerous numbers of the underlying perfect triads of IRBs, which form so-called universal numerical information field [5]. It is, essentially, a fundamental property of the space-time. The mutual connection between symmetry and non-symmetry of the IRBs provides a better understanding of physics of space-time, and role of geometric structure in the behaviour of natural and man-made objects.

7. IDEAL RING BUNDLE AS SYMMETRIC OBJECT

Next, we regard the n-stage ring sequence $K_{2D}=\{(k_{11},k_{12}),(k_{21},k_{22})...(k_{n1},k_{n2})\}$, where we require

all terms in each circular vector-sum to be consecutive 2-stage sequences as elements of the sequence. A circular vector-sum of consecutive terms in the ring sequence can have any of the n terms as its starting point, and can be of any length from 1 to n-1. An n-stage ring sequence K_{2D}, for which the set of all circular vector-sum forms two-dimensional grid, where each node of the grid occurs exactly R-times, is named a two-dimensional Ideal Ring Bundle (2-D IRB).

Here is two-dimensional n-stage ring sequence IRB with four (n=4) terms in the ring topology, where k_1= (0,2), k_2= (1,0), k_3=(1,1), k_4=(2,2), which grid chart is depicted below (Fig.7).

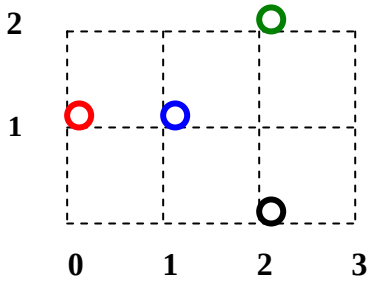


Fig. 7. Grid chart of two-dimensional Ideal Ring Bundle (2-D IRB) of four (n=4) vectors {(0, 2), (1, 0), (1, 1), (2, 2)}

We can calculate easy the all circular two-dimensional vector-sums, taking modulo m_1 =3 for the first component of vector-sum and modulo m_2 =4 for the second its component:

(0, 0) ≡ (2, 2) + (0, 2) + (1, 0),
(0, 1) ≡ (1, 1) + (2, 2) + (0, 2),
(0, 2) = (0, 2),
(0, 3) ≡ (1, 1) + (2, 2),
(1, 0) = (1, 0),
(1, 1) = (1, 1),
(1, 2) ≡ (0, 2) + (1, 0),
(1, 3) ≡ (1, 0) + (1, 1) + (2, 2),
(2, 0) ≡ (2, 2) + (0, 2),
(2, 1) ≡ (1, 0) + (1, 1),
(2, 2) = (2, 2),
(2, 3) ≡ (0, 2) + (1, 0) + (1, 1)

So long as the vectors (0, 2), (1, 0), (1, 1), (2, 2) of the ring sequence themselves are circular vector-sums too, the circular vector-sums set configure the 3×4 matrix as follows:

(0, 0)	(0, 1)	(0, 2)	(0, 3)
(1, 0)	(1, 1)	(1, 2)	(1, 3)
(2, 0)	(2, 1)	(2, 2)	(2, 3)

The result of the calculation forms the 3×4 grid which exhausts the circular 2-D vector-sums and each of its meets exactly once (R=1). So, the ring sequence of the 2-D vectors {(0, 2), (1, 0), (1, 1), (2, 2)} is two-dimensional Ideal Ring Bundle (2-D IRB) with n=4, R=1, and m_1 =3, m_2 =4.

It is easy to see, that 2-D IRB {(0, 2), (1, 0), (1, 1), (2, 2)} has the mirror symmetry, as well as differences of coordinates for all pairs of elements (bubbles on the grid chart) form a set of non-identical positions of 2-D vectors. So, the 2-D IRB is both symmetric and non-redundant structure at the same time. There are lots of symmetric and non-redundant 2-D IRBs. It is known, we can configure numerous numbers of one- and multidimensional IRBs [4].

As we see a whole can be partitioned into two parts using corresponding couple of rays, where one part of two partners of an arbitrary polarity is designed with the sign "+" and the other part of this polarity is designed with the sign "-", so in the following we shall always speak of the IRB polarity.

Perfect Distribution Law (PDL), based on the remarkable properties and structural perfection of one- and multidimensional IRBs are as follows:

A whole (space-time) can be partitioned perfectly, namely the sums of connected parts of the whole enumerate the set of harmony rising its parts exactly R-times with the smallest possible number of intersections.

The main characteristics of the PDL [5]:

1. A whole can be partitioned perfectly by infinitely great number of ways.
2. The minimal share of a whole in perfect partitioning can be anyone small.
3. The perfect partitioning can be useful to objects of anyone vector space dimensionality.

8. PERFECT VECTOR DATA CODING DESIGN

The development of new direction in fundamental and applied research in information technologies is connected with vector data coding design, based on the concept of Ideal Ring Bundles. For example, underlying combinatorial construction can be represented as mathematical model of optimum coding system, based on so-called "Monolithic Binary Code" (MBC). This code forms binary code combinations which all symbols "1" as well as "0" are arranged together. For example, two-dimensional vector data MBC formed on the 2-D IRB {(0,2), (1,0),(1,1),(2,2)}, is given in the Table 3.

Table 3. 2-D MBC based on the irb {(0,2), (1,0), (1,1), (2,2)}

Vector	Code	Vector	Code	Vector	Code
(0,0)	1101	(1,0)	0100	(2,0)	1001
(0,1)	1011	(1,1)	0010	(2,1)	0110
(0,2)	1000	(1,2)	1100	(2,2)	0001
(0,3)	0011	(1,3)	0111	(2,3)	1110

Table 3 contains set of binary code combinations for coding of all 2-D vectors on the 3×4 – matrix from (0,0) (code combination 1101) to (2,3) (code combination 1110), where each of them has been coded in circular 2-D MBC. The remarkable properties of the MBC provide its some advantages over the rest codes, involving simplicity of error detecting and correcting as well as high-speed operation.

8. CONCLUSION

The Perfect Distribution Law (PDL) discovers, essentially, a new scientific conception for development fundamental and applied research in information technologies, based on the idea of "perfect" combinatorial constructions- Ideal Ring Bundles (IRB)s, and the development of new directions in fundamental and applied research in area of information technologies, for improving such quality indices as reliability, precision, speed, resolving ability, and functionality. Structural perfection and remarkable properties of one- and multidimensional IRBs provide an ability to reproduce the maximum number of combinatorial varieties in the systems with a limited number of elements and bonds. It was IRBs provide many opportunities to apply them to advanced information technologies. The innovative methodologies based on combinatorial techniques of the Ideal Ring Bundles theory can be used for finding optimal solutions for wide classes of technological problems. Perspective information technologies and coded design prospected from the PDL are for example high performance vector data coding and self-checking code, vector computing machinery and low side-lobe antenna design [4]. The mutual connection between symmetry and non-symmetry of the IRBs provides a better understanding of physics of space-time, and role of geometric structure in the behaviour of natural and man-made objects. A perfection, beauty and harmony exists not only in the abstract models but in the real world also.

9. REFERENCES

[1] P.Wigner, Symmetries and reflections. Bloomington-London: Indiana University Press, 1970.

[2] H. Weyl, *Symmetry***.** Princeton, NJ: Princeton University Press, 1952.

[3] V. Riznyk. Combinatorial Sequencing Theory and its Applications, Lehrstuhl für Informatik V, Universität Mannheim (report), Germany 10 October 1997.

[4] S.Golomb, P.Osmera and V.Riznyk. Combinatorial Sequencing Theory for optimization of signal data vectors converting and signal processing. *Proceedings of the Workshop on Design Methodologies for Signal Processing*, Zakopane, Poland 29-30 August 1996, pp.43-44.

[5] V.Riznyk.Perspective information technologies prospected from basic laws of the world-wide harmony. *Proceedings of the Third International Conference on Computer Science and Information Technologies (CSIT'2008)***,** Lviv, Ukraine 25-27 September 2008, pp.80-84.

Volodymyr V. Riznyk *born in Polis'ke, Kyiv Region, Ukraine, 1940. Education: Diploma, Power Engineering, 1962; Diploma, Radio Engineering, 1967; PhD Degree, Cybernetics and Theory of Information, Physical and Mechanical Institute, Academy of Sciences, Ukr. SSR, 1980; D Sc, Mathematical Modeling and Mathematical Techniques for Scientific Researches, Vinnytsia State Technical University, 1994. Full professor in Lviv Polytechnic National University, 1995, University of Technology and Life Sciences Bydgoszcz, Poland, 1996.*

Scientific interests: researches and applications of combinatorial models and techniques to problems of information technologies.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

РОЗРАХУНОК РЕКРЕАЦІЙНОЇ ПРИВАБЛИВОСТІ ТЕРИТОРІЙ З ВИКОРИСТАННЯМ НЕЧІТКОЇ ЛОГІКИ

Я. Виклюк [1)], О. Артеменко [2)]

[1)] Національний університет "Львівська політехніка", вул. С.Бандери 12, Львів, 79013,
e-mail: vyklyuk@ukr.net
[2)] Буковинський університет, вул. В.Сімовича, 21, Чернівці, 58000
e-mail: o_hapon@yahoo.com

Резюме: *запропоновано метод розрахунку агрегованого потенціалу туристичної привабливості території на базі нечіткої логіки з врахуванням фактору сезонності. Визначено потенціали туристичної привабливості основних туристично-рекреаційних систем Чернівецької області.*

Ключові слова: *сезонна привабливість території, нечітка логіка, програмно-алгоритмічна модель.*

1. ВСТУП

Протягом XX сторіччя поряд з традиційними галузями економіки сформувалась і розвинулась туристична індустрія. Бізнес заснований на розвагах та організації різноманітного відпочинку надає підприємцям широкі можливості для отримання прибутків.

Різноманітний малий і середній туристичний бізнес швидко розвивається в багатьох регіонах країни і в Чернівецькій області зокрема. Побудовані та організовані без відповідного наукового обґрунтування туристичні комплекси, фірми, турбази, готелі тощо не отримують достатньої кількості замовлень, а отже виявляються нерентабельними.

Туристична галузь будь-якого регіону розвивалася б значно ефективніше, якби можна було визначати потенційно привабливі для туристів та відпочиваючих території, визначати рівень їх привабливості та спеціалізацію на відповідних видах відпочинку. Це дозволить виявити привабливі для інвестицій об'єкти, а також допоможе формувати більш ефективну стратегію економічного розвитку туризму в регіонах.

2. МЕТА ТА АКТУАЛЬНІСТЬ ДОСЛІДЖЕННЯ

Метою дослідження є розробка нечіткого алгоритму розрахунку потенційної туристичної привабливості території[1].

Актуальність дослідження у визначенні рівня привабливості території для туристів та відпочиваючих протягом року з метою формування стратегії діяльності підприємств туристичної та рекреаційної галузей.

Практична цінність статті полягає в наданні конкретних рекомендацій інвесторам, щодо доцільності створення туристично-рекреаційних систем (ТРС), та визначенні оптимальної стратегії їх діяльності.

3. ПОСТАНОВКА ЗАДАЧІ

Однією з проблем, що виникають при організації туристично-рекреаційного комплексу є сезонність його діяльності. Наприклад, гірськолижні комплекси отримують багато замовлень в зимовий період, але незаповнені решту року. Це пов'язано з відповідними кліматичними умовами. Тому, іноді є невигідним створення такого бізнесу, навіть тоді, коли територія має досить сприятливі для цього умови.

Деякі території мають сприятливі кліматичні, природні, матеріальні та інші умови для організації кількох видів відпочинку та рекреації, як в межах одного сезону так і протягом року. В зимовий період турбаза може працювати в режимі гірськолижного курорту, а, наприклад, влітку організувати розваги на воді, якщо поряд є відповідна водойма. Багатопрофільність туристичного підприємства не тільки збільшує його прибутки, але й робить менш залежним від несприятливих факторів тимчасового характеру, таких як невідповідні погодні умови протягом

тривалого періоду. Адже, якщо внаслідок цього недоотримані доходи в один з сезонів, є можливість компенсувати це активною діяльністю решту року. Чим більше джерел доходу у підприємства, тим воно стійкіше до впливу несприятливих факторів та форс-мажорних обставин. Тому, при проведенні відповідних аналітичних розрахунків, потрібно враховувати на вплив такого фактора, як сезонність.

Дане дослідження спрямоване на визначення сезонного агрегованого показника привабливості території для туристів та відпочиваючих з врахуванням різноманітності їх вподобань щодо відпочинку. Крім того, варто врахувати величину цільової аудиторії, тобто кількість людей, зацікавлених у даному виді відпочинку.

При розробці алгоритму визначення комплексного сезонного потенціалу туристичної привабливості території, найперше, потрібно обрати такий метод моделювання, що дозволить не тільки врахувати всі вищезгадані фактори, але й відобразить їх вплив якомога адекватніше.

При побудові складних математичних та формальних моделей виникає проблема рівня їх адекватності реальним умовам, особливо, якщо на прийняття рішення впливають якісні фактори. Такі фактори важко, а іноді й неможливо описати з допомогою класичного математичного інструментарію. Крім того, класичні методи побудови моделей не показують задовільних результатів, коли вхідні дані для опису та постановки задачі є апріорі неточними або неповними[2]. Тому для розв'язання нашої задачі не можливо створити повну та точну модель класичними методами. Щоб отримати адекватні результати необхідно підібрати такий математичний апарат, що дозволить оперувати неповними та якісними характеристиками.

Навіть у тих випадках, коли доступна лише частина потрібної інформації або відомості є досить розмитими, інтелект людини дозволяє їй приймати правильні рішення.

Науковцями розроблено та постійно вдосконалюється математичний апарат, який певною мірою повторює можливості людського інтелекту – теорію нечітких множин та нечітку логіку. Створене на їх базі нечітке моделювання є одним з провідних напрямків у прикладних та наукових дослідженнях. Нечітке моделювання є ефективним, коли в описі технічних систем чи бізнес-процесів присутня невизначеність, яка ускладнює або навіть унеможливлює застосування точних кількісних методів та підходів[3].

Нечітка логіка та нечітке моделювання зараз досить успішно застосовуються, зокрема, для розв'язання економічних задач. Процеси в деяких з цих задач раніше практично неможливо було описати або змоделювати. А створені моделі не давали повної картини ситуації, оскільки не враховували якісних факторів.

Нечіткі моделі виявились простішими та більш ефективними за класичні, зокрема, при оцінюванні глобального економічного рівня держави[4]. А такий показник як якість функціонування підприємства, взагалі не обчислювався математично без застосування нечіткої логіки, оскільки повинен враховувати багато факторів, що вимірюються різними величинами, а крім того, серед них багато якісних характеристик[5].

Останні дослідження, зокрема, китайських науковців показали зручність та ефективність використання нечіткого моделювання для розв'язання задач, подібних до даної[6].

3. МАТЕМАТИЧНА МОДЕЛЬ

Рекреаційна привабливість території визначається видами відпочинку та рекреації, які можна організувати та здійснювати на даній території. Відпочинок та рекреація, в свою чергу, залежать від кліматичних, географічних, історико-культурних умов та діяльності людини. Розрахунок рекреаційної привабливості буде проводитись для територій Чернівецької області. Тому для визначення цього показника модель включатиме параметри, що базуються на видах відпочинку, які під впливом вищезгаданих умов є актуальними для даних туристично-рекреаційних об'єктів (ТРО).

Отже, агрегований показник привабливості території для туристів та відпочиваючих складається з кількох окремих показників привабливості, що базуються на певних видах відпочинку. Для територій Чернівецької області актуальні види відпочинку та рекреації можна об'єднати в чотири групи:

p_1 – зимовий відпочинок;

p_2 – відпочинок в літній період на воді;

p_3 – відпочинок на природі весною-восени;

p_4 – екскурсії та огляд історико-культурних пам'яток.

Відповідно, сезонний рекреаційний потенціал території визначається як:

$$P(t) = f\left(p_1(t),\ldots,p_4(t)\right). \qquad (1)$$

Для обчислення агрегованого показника рекреаційної привабливості нами запропоновано скористатись лінійною згорткою, яка дозволяє

отримати інтегральний показник в тих випадках, коли вхідними змінними є незалежні та рівноцінні величини[7]:

$$P(t) = \sum_{i=1}^{4} p_i(t) \cdot \omega_i(t), \qquad (2)$$

де $\omega_i(t)$ – нормовані значення параметрів групових показників атрактивності.

Нормоване значення параметра ω_i розраховується за формулою:

$$\omega_i(t) = \frac{\omega_i^*(t)}{\sum_{i=1}^{n} \omega_i^*(t)}, \qquad (3)$$

де n – це загальна кількість параметрів даного потенціалу привабливості, а ω_i^* визначається як:

$$\omega_i(t)^* = C_i \cdot H_i(t), \qquad (4)$$

де C_i – відсоток людей, що бажають *i*-того виду відпочинку, $H_i(t)$ – сезонна можливість відпочинку.

Організація відпочинку в літній період року залежить від 7 основних параметрів, для яких визначені наступні лінгвістичні змінні:

x_1 – плавання;
x_2 – сплав на рафтах, байдарках та ін.;
x_3 – риболовля;
x_4 – катання на човнах, катамаранах тощо;
x_5 – тип водойми;
x_6 – якість під'їзних шляхів;
x_7 – підготовленість території для відпочинку.

У нашій моделі присутні лише три типи водойм: річка, озеро та ставок, оскільки тільки такі водойми присутні на території Чернівецької області. Відповідно, й відпочинок береться до уваги лише такий, який можна організувати на даних водоймах.

При створенні нечітких експертних систем, найбільш якісними є бази знань, у яких кількість вхідних параметрів не перевищує п'ять. Велика кількість вхідних параметрів значно ускладнює для експерта задачу опису причинно-наслідкових зв'язків з допомогою нечітких правил. Тому, при наявності великої кількості вхідних параметрів, їх потрібно ієрархічно класифікувати[8].

Ієрархічними є системи нечіткого виводу, в яких вивід однієї бази знань подається як вхідний параметр іншої, що знаходиться на вищому рівні ієрархії. В таких системах відсутні зворотні зв'язки. Ієрархічні системи нечіткого виводу використовуються при моделюванні складних систем з багатомірними залежностями "вхід – вихід".

Однією з переваг ієрархічних систем є компактність баз знань у підсистемах. Зв'язки в такій базі знань можна адекватно описати невеликою кількістю продукційних правил, при чому це будуть короткі правила з двома-трьома вхідними змінними. При побудові нечіткого виводу в ієрархічній системі не виконуються процедури дефаззифікації та фаззифікації для проміжних змінних. Результат логічного виводу однієї підсистеми одразу подається у вигляді нечіткої множини на вхід підсистеми вищого ступеня ієрархії.

Нами запропоновано для обчислення рекреаційного потенціалу літнього відпочинку створити дві підсистеми. Перша об'єднує підвиди відпочинку на воді та визначає потенційну кількість видів відпочинку, доступних для даної водойми:

$$p_{11} = f(x_1, ..., x_4). \qquad (5)$$

Цей показник є одним з вхідних параметрів іншої підсистеми, яка і визначає сумарний потенціал літнього відпочинку на даній території:

$$p_1 = f(p_{11}, x_5, x_6, x_7). \qquad (6)$$

Зимовий відпочинок в основному пов'язаний з гірськолижними видами. Це, зокрема, актуально для Чернівецької області, рельєф якої є переважно гірським. За останні кілька років підприємці відкрили більше 10 гірськолижних баз в різних районах області.

На думку експертів рівень сприятливості умов для організації та ведення туристичного бізнесу в напрямку гірськолижного відпочинку впливають такі фактори:

x_8 – висота схилу;
x_9 – довжина схилу;
x_{10} – експозиція схилу;
x_{11} – якість під'їзних шляхів.

Від того, наскільки високо над рівнем моря розташовано вершину схилу, залежить тривалість існування снігового покриву на ньому, а отже, тривалість сезону. Також на тривалість сезону впливає положення схилу, оскільки північний схил менше прогрівається сонячними променями. Відповідно, сніг тане не так швидко, як на південному схилі. Довжина схилу визначає довжину та кількість маршрутів, які можна прокласти. Крім того, чим довший

схил, тим більше відпочиваючих можуть там перебувати одночасно. Якість під'їзних шляхів є одним з визначальних факторів для тих туристів, які дістаються до місця відпочинку власним автомобілем. Тому потенціал привабливості території для зимового відпочинку визначається:

$$p_2 = f(x_8,...,x_{11}). \quad (7)$$

Відпочинок весною та восени в основному полягає в проведенні вихідних днів на природі. Як правило, туристи не віддаляються від свого дому на значні відстані. Основними чинниками, що впливають на потенційну привабливість даної території для туристів весною та восени є можливість отримати туристичні послуги. Найпопулярнішими у вказаний період року є:

x_{12} – проведення пікніків;

x_{13} – збір ягід, грибів та іншого;

x_{14} – інші розваги на природі (наприклад, катання на конях, велосипедах тощо).

Отже, груповий показник атрактивності відпочинку весною та восени розраховується як:

$$p_3 = f(x_{12}, x_{13}, x_{14}). \quad (8)$$

Останній груповий показник – потенціал історико-культурної привабливості об'єкта був розрахований в роботі[9]. Цей показник залежить від двох параметрів:

- географічних координат історико-культурних пам'яток та цікавих для туристів місць Чернівецької області;
- рейтингових оцінок значимості вищевказаних об'єктів (визначаються експертами).

Показник історико-культурної атрактивності території визначається:

$$p_4' = \sum_{i=1}^{m} \left(\pi_i \times e^{-\frac{(N-\pi_i) r_{kl,i}^2}{\sigma^2}} \right), \quad (9)$$

де: $r_{kl,i}$ – відстань між територією, для якої обчислюється потенціал та ТРО, що має історико-культурне значення; σ – середньоквадратичне відхилення визначає форму функції, (квантель порядку ½ визначає "оптимальну відстань" при якій потенціал ТРО спадає вдвічі); π – рейтингова оцінка рекреаційного потенціалу історико-культурного ТРО; N – максимальне значення рейтингу (при $m=N$ всі відвідувачі відвідають рекреаційний об'єкт).

Коефіцієнт історико-культурної привабливості території показує, наскільки оптимально розташовано даний туристичний об'єкт відносно основних історико-культурних пам'яток Чернівецької області, тобто тих місць, що є цікавими для огляду туристами. Наявність поблизу таких об'єктів дозволяє організовувати екскурсійні поїздки, що може урізноманітнити перелік пропонованих послуг. Відповідно, чим більше цікавих для туристів місць є досяжними, тим більше значення показника потенціалу історико-культурної привабливості.

Значення коефіцієнта історико-культурної привабливості (9) для різних ТРО може сильно відрізнятись в розрядності числа, що негативно вплине на основний результат. Тому для потенціалу істотко-культурної привабливості виконується нормування з допомогою нечіткого алгоритму, в якому цей показник представлено у вигляді лінгвістичної змінної. В результаті нормування значення коефіцієнта розташовуються в діапазоні від 0 до 1.

Схематично модель розрахунку агрегованого показника рекреаційної привабливості території зображено на Рис. 1.

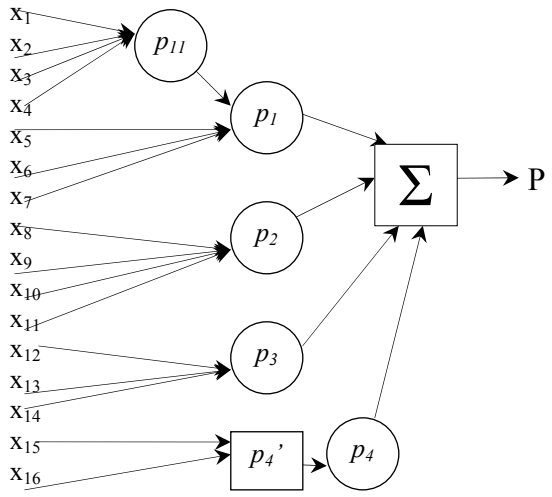


Рис.1 – Модель визначення комплексного сезонного потенціалу туристичної привабливості території

Отже, групові потенціали визначено за основними видами відпочинку, які можна проводити на території Чернівецької області протягом року. Кожен з групових показників привабливості залежить від кількох факторів, які впливають на можливість та якість організації даного виду відпочинку. Виділено групи факторів: зимового відпочинку, відпочинку на воді, відпочинку в весняно-осінній період та літнього відпочинку. Для кожної групи розраховується потенціал привабливості

території для туристів та відпочиваючих.

Загалом, комплексний сезонний потенціал привабливості території для відпочиваючих та туристів залежить від 16 основних вхідних параметрів, 14 з яких представлено у вигляді нечітких лінгвістичних змінних.

4. КОМП'ЮТЕРНИЙ ЕКСПЕРИМЕНТ

Для визначення термів лінгвістичних змінних $x_1,\ldots,x_{14}$ та коефіцієнта історико-культурної привабливості ми використовували трикутні, трапецієвидні, S-подібні та Z-подібні функції приналежності. Функція приналежності задається параметрами, кількість яких, залежно від виду функції, становить від двох до чотирьох[2].

Параметри функцій приналежності, визначені для лінгвістичних змінних, що використані в даній моделі, отримано згідно даних відділу з питань туризму Чернівецької облдерж-адміністрації.

Сезонна можливість здійснення відпочинку $H(t)$ визначено за допомогою експертних оцінок та приведено в таблиці 1.

Таблиця 1. Сезонна можливість здійснення відпочинку даного виду

Місяці (t)	Види відпочинку			
	Зимовий	Літній	Історико-культурний	Весною-восени
Січень	1	0	0,1	0
Лютий	0,9	0	0,1	0
Березень	0,5	0,1	0,6	0,4
Квітень	0	0,3	0,8	0,9
Травень	0	0,5	1	1
Червень	0	1	0,6	0,5
Липень	0	1	0,6	0,5
Серпень	0	1	0,7	0,7
Вересень	0	0,5	0,8	0,8
Жовтень	0	0,4	1	1
Листопад	0,3	0	0,5	0,7
Грудень	1	0	0,1	0

Кількість туристів, зацікавлених певним видом відпочинку у відсотках до всієї маси потенційних туристів C_i, визначено маркетинговими дослідженнями, його значення для різних груп показано у таблиці 2[10].

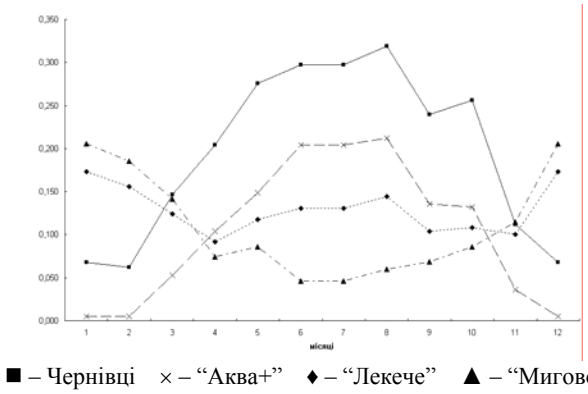
Таблиця 2. Частина потенційних споживачів певного відпочинку серед усієї маси туристів.

Вид відпочинку	Відсоток туристів, бажаючих споживати, %
Зимовий	50
Літній	90
Історико-культурний	65
Весною-восени	80

В нашій роботі розраховувались показники комплексних сезонних потенціалів привабливості для 10 туристичних об'єктів Чернівецької області. Дані об'єкти є популярними центрами відпочинку, розташованими в різних частинах області. Всі об'єкти відрізняються за природними умовами, рельєфом, розмірами та рівнем розвитку інфраструктури.

За результатами комп'ютерних розрахунків можна прослідкувати динаміку зміни комплексної рекреаційної привабливості території протягом року(Рис.2).

На рисунку показано помісячну динаміку комплексного показника рекреаційної привабливості чотирьох популярних туристично-рекреаційних об'єктів Чернівецької області: місто Чернівці, розважальний комплекс "Аква+", садиба зеленого туризму "Лекече" та гірськолижний комплекс "Мигово".



■ – Чернівці × – "Аква+" ♦ – "Лекече" ▲ – "Мигово"

Рис.2 – Зміна потенціалу рекреаційної привабливості деяких туристично-рекреаційних об'єктів Чернівецької області

Чернівці є великим ТРО, що надає різноманітні туристичні послуги. Рівень привабливості цієї території є досить високим протягом всього року. Зниження значення агрегованого показника в зимовий період пов'язане з тим, що дана територія взимку є привабливою в основному за рахунок гірськолижної траси на горі Цецино. Тоді як в інші місяці туристів приваблюють і історико-культурні об'єкти і можливість покататись на

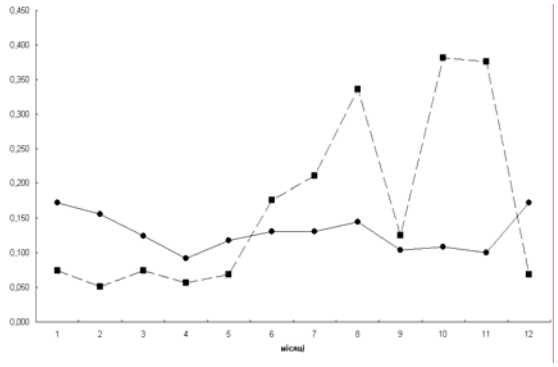
конях (Цецино), провести пікнік (Кемпінг), а також відпочинок коло водойми.

Розважальний комплекс "Аква+" розміщено на березі озера. Основною спеціалізацією цього об'єкта є послуги організації відпочинку на воді, а також проведення пікніків тощо. З рисунку видно, що пік привабливості даного ТРО припадає на літні місяці. Весною та восени показник атрактивності зменшується, а в зимовий період – близький до нуля.

Гірськолижний комплекс "Мигово" навпаки, має високі значення агрегованого показника рекреаційної привабливості в зимовий період, але є нецікавим для туристів та відпочиваючих решту року.

Садиба зеленого туризму "Лекече" є невеликим ТРО, що надає різноманітні туристичні послуги: сплав на рафтах і катамаранах, полювання, збір грибів, ягід, катання на конях та ін. Крім того, дана територія має сприятливі умови для організації гірськолижного відпочинку. Рівень рекреаційної привабливості даної території є досить високим впродовж всього року. На відміну від попередніх об'єктів, на графіку "Лекече" відсутні різкі піки та спади.

Динаміка потенціалу рекреаційної атрактивності для вищезгаданого ТРО в порівнянні з відносною кількістю відпочиваючих наведено на Рис.3. З рисунку видно, що отримана залежність туристичної привабливості об'єкта від сезону добре корелює з експериментальними даними.



● – розрахунок ■ – експеримент

Рис. 3 – Сезонна залежність комплексного потенціалу привабливості в порівнянні із відносною кількістю рекреантів, на прикладі садиби "Лекече"

Крім того, можна відмітити, що існує певна функціональна залежність між рівнем рекреаційного потенціалу та кількістю клієнтів.

Показники комплексного рекреаційного потенціалу для даного об'єкта в червні та липні мають однакові значення, тобто умови для відпочинку у ці місяці є однаково сприятливими. А статистика показує, що в липні відпочиваючих було більше, ніж в червні. Розбіжність в показниках пояснюється тим, що червень є лише початком сезону відпусток та канікул. У липні ж більше людей їдуть відпочивати. При плануванні відпочинку люди зважають також на те, що початок літа може бути прохолодним, тому щоб напевне добре відпочити йдуть у відпустки не раніше як в кінці червня. Значний спад кількості відпочиваючих у вересні пов'язаний не лише зі зниженням рекреаційної привабливості даного ТРО, але й завершенням сезону відпусток.

Збільшення кількості відпочиваючих починаючи з серпня є результатом проведення рекламної кампанії в травні-червні. Протягом цих місяців було розміщено оголошення в періодичних виданнях Чернівецької області та туристичних путівниках.

Розбіжність в динаміці між розрахунковими та статистичними даними в зимовий період (грудень, січень) зумовлена тим, що даний туристично-рекреаційний об'єкт у своїх рекламних акціях не позиціонує себе як організатора гірськолижного відпочинку. Хоча й має досить сприятливі умови для організації та проведення такого відпочинку.

Використання узагальненого показника дає змогу більш адекватно визначити привабливість території для туристів та відпочиваючих, а отже, показує потенційним інвесторам та відповідним органам місцевої влади реальні перспективи використання рекреаційних ресурсів даного району.

5. ВИСНОВКИ

В роботі запропоновано метод розрахунку комплексного потенціалу туристичної привабливості території на базі нечіткої логіки. Даний метод дозволяє врахувати наявні природні умови та інфраструктуру для організації та проведення різноманітних видів відпочинку та розважальних заходів.

Вперше при побудові потенціалу привабливості території враховано фактор сезонності.

Метод розрахунку агрегованого показника туристичної привабливості території апробовано для десяти популярних туристично-рекреаційних об'єктів Чернівецької області.

Запропонований метод дозволить підприємствам туристичної галузі більш адекватно обирати напрям та масштаби капіталовкладень при плануванні стратегії діяльності, проведенні PR-акцій.

В поєднані з ГІС технологіями розроблений метод дозволить отримати карту потенційної

туристичної привабливості території, що може слугувати науковим підгрунтям стратегії розвитку регіону. Запропонований метод дозволить в подальшому включити до алгоритму сегментацію цільової аудиторії – відпочиваючих залежно від їх вподобань та фінансових можливостей.

6. СПИСОК ЛІТЕРАТУРИ

[1] Виклюк Я.І. Побудова fuzzy-моделі для визначення рекреаційного потенціалу євро регіону "Верхній Прут" / Я.І. Виклюк // Вестник НТУ "ХПИ". Сборник научных трудов. Тематический выпуск "Системный анализ, управление и информационные технологии". Харьков: НТУ "ХПИ", 2007. – №41, с. 193–201.

[2] Леоненков А.В. Нечеткое моделирование в среде MATLAB и fuzzyTECH / А.В. Леоненков; – СПб.: БХВ-Петербург, 2005. – 736 с.

[3] Дьяконов В.П. MATLAB 6.5 SP1/7/7 SP1/7 SP2 Simulink 5/6 Инструменты исскуственного интеллекта и биоинформатики / В.П.Дьяконов, В.П.Круглов – Серия "Библиотека профессионала". – М.:СОЛОН-ПРЕСС, 2006.–456с.

[4] Иманов К.Д. Нечеткая модель определения метаэкономического уровня / К.Д. Иманов, Р.Р. Рзаев // Системні дослідження та інформаційні технології, 2006, №4.

[5] Ткачук Л. М. Економіко-математичне моделювання якості функціонування підприємства / Л. М. Ткачук // Інформаційні технології та комп'ютерна інженерія, 2006, №1(5).

[6] Shengquan Ma Fuzzy model of regional economic competitiveness in GIS spatial analysis: Case study of Gansu,Western China /, Jing Feng, Huhua Cao // Fuzzy Optim Decis Making, 2006. – №5, p.99–111.

[7] Гнатієнко Г. М. Експертні технології прийняття рішень: Монографія. / Г.М. Гнатієнко, В.Є. Снитюк – К.: ТОВ "Маклаут", – 2008. – 444 с.

[8] Штовба С. Д. Проектирование нечетких систем средствами MATLAB. / С. Д. Штовба – М.: Горячая линия – Телеком, 2007. – 288 с.

[9] Проблеми географії та менеджменту туризму:Монографія. / [Якін В.Г., Руденко В.П., Король О.Д. та ін.] – Чернівці: Чернівецький національний університет імені Юрія Федьковича, 2006.

[10] Дурович А.П. Маркетинг в туризме: Учеб. пособие / А.П. Дурович. – 3-е изд., стереотип. – Мн.: Новое знание, 2003. – 496 с.

***Ярослав Виклюк**, докторант НУ "Львівська політехніка", доцент, к.ф.-м.н. В 1999 році закінчив фізичний факультет Чернівецького національного університету імені Юрія Федьковича за спеціальністю теоретична фізика. В 2002 році захистив кандидатську дисертацію за напрямком "фізика напівпровідників та діелектриків". В 2006 році отримав другу вищу освіту в Чернівецькому торговельно-економічному інституті КНТЕУ за напрямком економіка. Кандидат в майстри спорту зі спортивного туризму та спортивної гімнастики.*

Наукові інтереси: SoftComputing, еконофізика, інформаційні технології.

***Ольга Артеменко,** викладач кафедри Комп'ютерних систем і технологій Буковинського університету.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

CALCULATION OF THE TERRITORY RECREATION ATTRACTIVENESS USING FUZZY LOGIC

Yaroslav Vyklyuk [1)], Olga Artemenko [2)]

[1)] NU "L'vivska Politechnika",
Ukraine, 79013, L'viv, S.Bandera str.12,
e-mail: vyklyuk@ukr.net
[2)] Bucovinian University, Simovich str. 21, Chernivtsi, 58000, Ukraine
e-mail: o_hapon@yahoo.com

Abstract: *in this paper the calculation method of the territory tourist attractiveness aggregated potential is offered on the base of fuzzy logic taking into account the factor of seasonality. The potentials of tourist attractiveness are found for the basic tourist recreation systems of the Chernivtsi region.*

Keywords: *seasonal attractiveness of the territory, fuzzy logic, program-algorithmic model.*

Tourist industry of any region would develop considerably more effective, if it is possible to determine areas potentially attractive for tourists and holiday-makers, to determine the level of their attractiveness and specialization on the proper types of rest. It will allow finding out objects attractive for investments, and also will help to form more effective economic development strategy for tourist business in regions.

A research purpose is to develop fuzzy algorithm for calculation of the potential of territory's tourist attractiveness.

Research actuality is in determination of the level of territory's attractiveness for tourists and holiday-makers during a year with the purpose to form a strategy of activity for enterprises in tourist and recreation industries.

The practical value of this article consists in the giving concrete recommendations to the investors about expedience of creation the tourist recreation systems (TRS) and determination of the optimum strategy of their activity.

Among the problems which arise up during organization of tourist-recreation complex there is seasonality of its activity. For example, the mountain-skiing complexes get a lot of custom in winter, but are unfilled the rest of the year. It is related to the proper climatic terms. That is why, sometimes foundation of such business is unprofitable, even when territory has conditions complimentary enough for this purpose. Some territories have good climatic, natural, financial and other conditions for organization of a few types of rest and recreation, as within the limits of one season so for a year. In winter a tourist center can work as a mountain-skiing resort, and, for example, in summer to organize entertainments on water, if alongside there is the proper basin. Service diversification of tourist enterprise not only multiplies his incomes but also does less dependency upon the critical factors of temporal character, such as incongruous weather conditions during the long period of time. If, as a result of that, profits are received less in one of seasons, there is possibility to compensate it with active work of the rest of a year. The more sources of income has the enterprise, the more resistant it is to influencing of critical factors and circumstances of force-majeure.

The given research was directed on determination of the territory's seasonal aggregated index of attractiveness for tourists and holiday-makers conducting the variety of their preferences in having a rest. In addition, it should be taken into account the size of consumer audience, in other words, the amount of people, interested in this type of rest.

The recreation attractiveness of territory is determined by the types of rest and recreation, which can be organized and carried out on this territory. Rest and recreation, in the turn, depend on climatic, geographical, historic and cultural conditions and human activity.

Thus, the territory's aggregated index of attractiveness for tourists and holiday-makers consists of a few separate indexes of attractiveness,

which are based on the certain types of rest. For territories of the Chernivtsi region the actual types of rest and recreation can be united in four groups:

p_1 – winter rest;

p_2 – rest in summer time at basin;

p_3 – rest in spring and autumn in the countryside;

p_4 – excursions and review of historic and cultural sights.

The seasonal recreation potential of territory is determined as:

$$P(t) = f(p_1(t),..., p_4(t)). \quad (1)$$

The linear convolution product is used for the calculation of the aggregated recreation attractiveness index. It allows getting the integral index in those cases, when the input variables are independent and equivalent values:

$$P(t) = \sum_{i=1}^{4} p_i(t) \cdot \omega_i(t), \quad (2)$$

where $\omega_i(t)$ – are the parameters of groups attractiveness indexes normalized values.

The normalized value of coefficient ω_i is estimated with a formula:

$$\omega_i(t) = \frac{\omega_i^*(t)}{\sum_{i=1}^{n} \omega_i^*(t)}, \quad (3)$$

where n – is the common amount of parameters in given attractiveness potential, and ω_i^* is determined as:

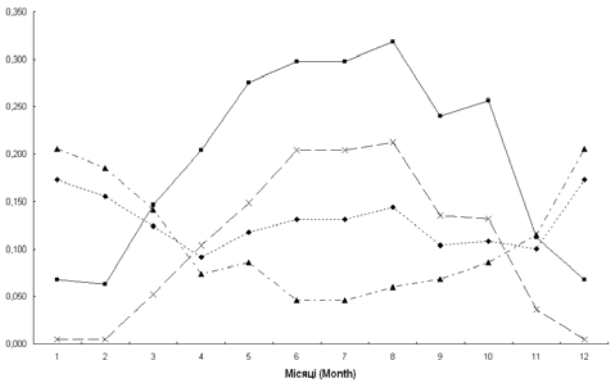
$$\omega_i(t)^* = C_i \cdot H_i(t), \quad (4)$$

where C_i – a percent of people that want to have the indicated type of rest, $H_i(t)$ – is seasonal possibility of having a rest.

In general, complex seasonal potential of territory's attractiveness for holiday-makers and tourists depends on 16 basic parameters, 14 of them are presented as fuzzy linguistic variables.

In this research the values of complex seasonal potentials of attractiveness were accounted for 10 tourist places in Chernivtsi region. These places are the popular centers for having a rest, located in different parts of region. All places differ in their natural conditions, relief, sizes and a level of infrastructure development.

As a result of computing it is possible to trace the variation dynamics of territory's complex recreation attractiveness during a year (fig. 1).



■ – Chernivtsi × – "Aqua+" ♦ – "Lekeche" ▲ – "Mygovo"

Fig. 1 – Variation of recreation attractiveness potential for some tourist-recreation centers of Chernivtsi region

Chernivtsi is a large tourist-recreation center (TRC) that provides various tourist services. This territory's level of attractiveness is high enough during the whole year. The decline of value of the aggregated index in winter is related to the fact that given territory is attractive in winter mainly due to a mountain-skiing route on the mountain of Tsetsino. While in other seasons tourists are also attracted with historic and cultural sights and possibility to ride horses (Tsetsino), to have a picnic (the Campsite), and also to take a rest near the river.

The entertaining complex "Aqua+" is placed ashore a lake. Basic specializations of this center are services of rest on water organization and also providing picnics and others like that. The picture describes that maximum level of this TRC attractiveness is summer period. In spring and in autumn the index of attractiveness decreases, and in winter period it falls down near to zero.

Mountain-skiing complex "Mygovo" opposite, has high values of the recreation attractiveness aggregated index in winter, but is uninteresting for tourists and holiday-makers the other seasons of the year.

The eco-tourism center "Lekeche" is a small TRC that provides various tourist services: rafting, fishing, hunting, gathering mushrooms, berries, horse riding and others. In addition, this territory has good conditions for organization of mountain-skiing rest. Level of recreation attractiveness of this territory is high enough during the year. Unlike previous places, on the graph "Lekeche" sharp increases and breakdowns are absent.

Using the aggregated index enables to define the attractiveness of territory for tourists and holiday-

makers more adequately, and consequently, shows to the potential investors and responsible institutions of local governing the real prospects of the recreation resources exploitation of this area.

In general, in this paper the method for calculation of complex tourist attractiveness potential of the territory is offered on the base of fuzzy logic. The developed method allows considering present natural conditions and infrastructure for organization and providing various types of rest and entertaining events.

For the first time in the calculation of territory's attractiveness potential the factor of seasonality is conducted.

The offered method will allow the tourist industry enterprises to elect more effectively direction and dimensions of capital investments in planning their strategy, arranging PR-actions.

Integrated to GIS technologies, the developed method will allow getting the map of potential tourist attractiveness of territory which can serve as a scientific base for strategy of regional development. The offered method will allow in future to plug in to the algorithm a segmentation of consumer audience – holiday-makers depending on their preferences and financial capabilities.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

SOFTWARE FOR DETERMINATION OF BEHAVIOR PATTERNS OF WEB-FORUM MEMBERS

Yuriy Syerov, Ruslan Kravets

National University "Lviv Polytechnics", 12 Bandery Street, Lviv, 79013 Ukraine
syerov@ridne.net, rkravets@ua.fm

Abstract: *article considers actual problem of Web-forums' members' behavior and classifying information system development. In the article existing Web-forum CMS' were analyzed, process of Web-community members' behavior analyzing was researched and overviewed, the process of information system developing was described.*

Keywords: *Web-community, forum, internet-forum, Web-community member, members' classes.*

INTRODUCTION

On the modern day level of Internet and global information system World Wide Web development, especially with appearance of new conception – Web 2.0, very important role is played by Web-communities, which consist of everyday Internet users. Web-forums are one of the most popular and effective, from the point of view of information representation and accessibility, forms of Web-community organization [1, 8].

Primary Web-forum's objects are members and content which they create: threads and posts. Web-forum member is a person that visits Web-forum site, reads, and publishes information in threads and posts. Web-forum content is a hierarchy of threads and posts, tree with two types of nodes: sub-forums and discussions.

Nowadays, most modern Web-forums function on the basis of popular software – Content Management System (CMS), for example: vBulletin, Invision, phpBB, XMB [1-3].

Research and analysis of Web-forums' members' behavior is a very complicated computational task. That is why development of information system of analysis of Web-forums' members' behavior, which should simply be integrated into modern CMS, is a very important task.

Members influence Web-forum's life that is why the Web-community developers need software, which lets them analyze members' behavior that helps to provide effective moderation and administration.

1. BACKGROUND

Web-forum functioning is impossible without a program complex, which provides basic tasks such as: member registration, content publishing, content, and member management. Such modern day complexes, such as CMS, have typical data structure scheme and provides all basic means that make it possible for Web-forums to function. The majority of Web-forums are based on popular CMS that have similar interfaces and functionality. This helps Web-forums members to familiarize themselves on new Web-forum quickly. Also, this makes it possible to develop informational system of analysis of Web-forums' members behavior for Web-forum administrators.

Modern CMS have good functionality, but they do not give administrators the means to analyze the Web-forum member's behavior. CMS provides only the simplest statistical information: post count, thread count, quantity of new members that registered during the last day, number of existing members that visited the Web-forum today, etc. The system does not give detailed and more complex information about content growth dynamics, member growth dynamics, or member's usefulness. That is why it is needed to develop a system that will provide the means of analyzing member behavior during certain periods of time, classifying members, provide decision support during Web-forum administration [1-3,8,9].

2. ARTICLE AIMS

The purpose of this article is to describe the process of information system of analysis of Web-

forums' members' behavior development.

Before we start system development it is necessary to research modern Web-forum CMS structures and the process of Web-forums' members' behavior analyzing. The reason this is necessary is because, our information system must solve all tasks of members' behavior analysis and be easily integrated into existing Web-forum CMS. On the basis of results of this research, we should plan a system structure and develop its prototype. So, the aims of this article are:

- analysis of the existing Web-forum CMS;
- research of the process of Web-forums member behavior analysis;
- development of data structure scheme of information system;
- program realization of algorithms of the Web-forums' members behavior analysis.

3. MAIN PART

3.1. STRUCTURAL AND INFORMATION WEB-FORUM MODEL

Primary Web-forums' objects are members and content, which they create: threads and posts.

Web-forum member is a person that visits Web-forum site, reads, and publishes information in threads and posts.

According to their abilities Web-forums' members belong to one of the following classes: unregistered visitors (guests), registered members, moderators, administrators.

Sub-forum is a set of lower-level sub-forums and threads.

Thread is a set of posts, created by Web-forum members. Threads could be created on any structure level, usually on the lowest level.

Web-forum content is a hierarchy of threads and posts (see Fig. 1).

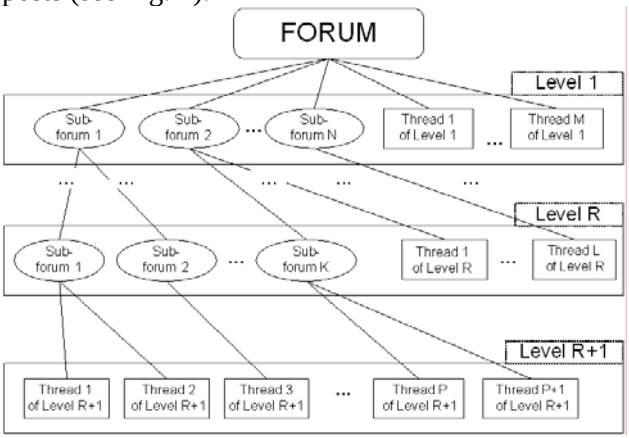


Fig. 1 – Tree-like Web-forum structure

Web-forum consists of a set of first-level forums. Each of them also consists of a set of second-level forums (sub-forums) and threads. Web-forum structure can be detailed more and more using lower levels. Web-forum structure could be changed during the Web-community's life. If members accumulate great amount of content more sub-forums on the lower levels could be created.

Forum – is a set of sub-forums and threads. Each sub-forum can consist of lower-level sub-forums and threads. In the process of creating forum structure it is recommended to use lower-level sub-forums. Using lower-level sub-forums helps to create better content semantic map.

Thread – is a set of posts, written by forum members. Threads could be placed on any level, but it is better if they are only at the lowest level. Main Web-forum feature is that threads are stored untill the forum exists for future members; in some rare specific situations when administrator takes decision to delete one of them.

Poll is a kind of discussion that aims to know member's opinion on some question. Poll start is a question with a set of answers with some explanation. Set of answers must be complete (all available answers). Member can write posts here like in any thread.

Post is a message written by a forum member.

Each post can be one of the following types:

- Thread start – first message in the discussion. It is very important to create correct understandable thread name.
- Answer-post. Each post in the thread that is not thread start. Answer-post may include previous parts of the previous posts – citations.
- Thread stop – last post in the thread, usually made by administrator or moderator when discussion is over or members start to write off-topic or flame posts.

3.2. GENERALIZED WEB-COMMUNITY INFORMATION MODEL

In general all Web-forum CMS have complex data scheme, which consists of great amount of tables. However, large part of these tables is secondary and store data that is not mandatory for Web-forum analysis.

After examining and analyzing existing CMS's, we can select only the main objects of such information system as Web-community. Generalized data scheme, which gives us all necessary information for Web-community member behavior analysis is shown on Fig. 2.

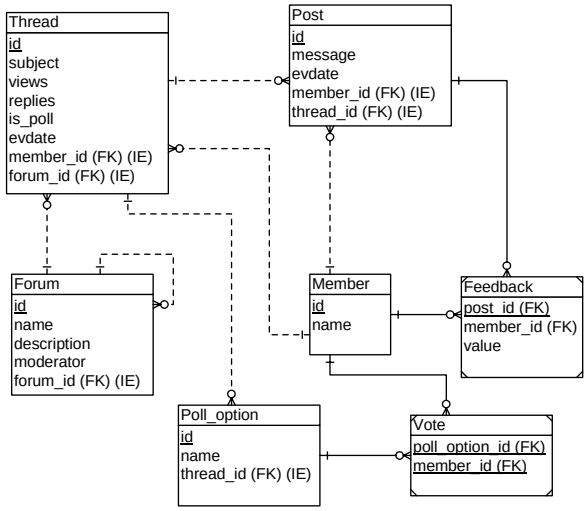


Fig. 2 – Generalized data scheme of web-forum CMS

Main parts of such information system, as Web-forum, are Members and content.

Content consists of Forums, Threads, and Posts.

A Web-forum's member is characterized by such parameters:

Name, Password, Personal Info, Set of created posts, Set of created threads, Set of received feedbacks, Set of given feedbacks.

Another part of this system – content, is presented by three entities: Forum, Thread, and Post.

Forum is characterized by such parameters:

Name, Number of threads (that belong to this forum), Number of posts (that belong to this forum), Description.

The thread is characterized by the following parameters:

Title, Author, Sub-forum (to which the thread belongs), Number of posts, Number of views.

The post is characterized by the following parameters:

Title, Body, Author, Data, Thread (to which the post belongs), Sub-forum (to which post belongs).

Prior to the system information model description, let us shortly overview Web-forums' members' behaviors, their features, and classification methods.

Web-community members may have the following features: Activeness, Creativeness, Attractiveness, Reactivity, and Loyalty.

For Web-community member features modeling, fuzzy sets are used [2]. The features of all listed members can be defined according to their activity within the forum. Activeness is defined by quantity of content that the members have created. Creativeness is defined by the quality of content that the members have created. Attractiveness is defined by the number of members who give feedback to the created content. Reactivity is defined by the way the member takes part in discussions. Loyalty is defined by the reaction to other member's content.

3.3. INFORMATION DATA MODEL FOR CLASSIFYING MEMBERS

The first step of creating information system of web community members behavior analyzing and classifying is development of data structure, mandatory for analysis. In order to create information system, it is necessary to develop metadata structure, which will describe the classes of members, rules of classification, and linguistic variables, which describe the features of members' behavior.

Developed information model looks as follows (Fig. 3.):

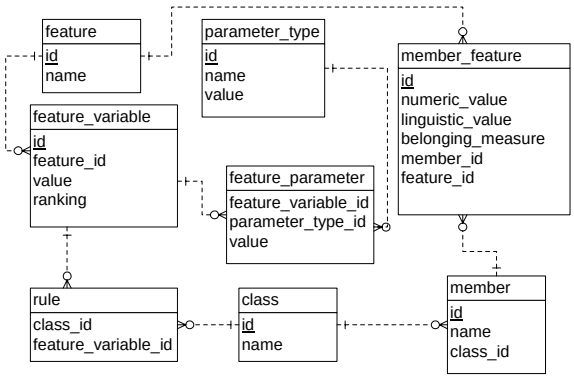


Fig. 3 – Data scheme of an information system of web community members behavior analyzing and classifying

Developed data scheme of information system of Web-forums' members' behavior analyzing and classifying will include metadata, mandatory for classification of Web-forum members. It integrates with the database of CMS Web-forum oh the basis of entity "Author" and consists of the following tables:

- Table *member* – keeps the information about all Web-community members;
- Table *class* – describes all Web-community members' classes;
- Table *rule* – describes the rules of Web-forum members' classification;
- Table *member_feature* – dedicated to keeping data about all members' characteristics;
- Table *feature* – describes all Web-community members' features;
- Table *parameter_type* – keeps the information about intervals of each linguistic variable.
- Table *feature_variable* – describes a domain of every linguistic variable – features of a Web-community member.
- Table *feature_parameter* – keeps the information about range of each linguistic variable.

3.4. CALCULATION OF MEMBERS' VALUE CHARACTERISTIC

Members' features will be calculated as follows.

Activeness of threads creation we will present as division of the number of threads, created by the member, by the total number of threads created by all the members:

$$Activeness_{Thread}(member_i) = \frac{card(Thread(member_i))}{card(Thread)},$$

where $card(X)$ – power of the set X;

$Thread(member_i)$ – the set of all threads, created by i-th member;

$Thread$ – set of all threads.

In the database this computation is realized with the following query:

```
SELECT
   member_id,
   COUNT(*)*100.0/(SELECT COUNT(*)
   FROM Thread) as ThreadActiveness
FROM Thread
GROUP BY member_id
```

Activeness of polls creation is calculated similarly to activeness of threads creation:

$$Activeness_{Poll}(member_i) = \frac{card(Poll(member_i))}{card(Poll)},$$

where $Poll(member_i)$ — the set of all polls, created by i-th member;

$Poll$ – set of all polls.

Query for activeness of polls creation looks as follows:

```
SELECT
   member_id,
   COUNT(*)*100.0/(SELECT COUNT(*)
   FROM Thread WHERE is_poll='y') as
   PollActiveness
FROM Thread
WHERE is_poll='y'
GROUP BY member_id
```

Activeness of posts creation we will present as division of the number of posts, created by the member, by the total number of posts created by all the members:

$$Activeness_{Post}(member_i) = \frac{card(Post(member_i))}{card(Post)},$$

where $Post(member_i)$ – the set of all posts, created by i-th member;

$Post$ – set of all posts.

Query for activeness of posts creation looks as follows:

```
SELECT
   member_id,
   COUNT(*)*100.0/(SELECT COUNT(*)
   FROM Post) as PostActiveness
FROM Post
GROUP BY member_id
```

Activeness of voting we will present as division of the number of polls, in which the member took part, by the total number of polls created by all the members.

$$Activeness_{Vote}(member_i) = \frac{card(Vote(member_i))}{card(Poll)},$$

$Vote(member_i)$ – all polls in which i-th member took part;

$Poll$ – set of all polls.

Query for activeness of voting looks as follows:

```
SELECT
   member_id,
   COUNT(*)*100.0/(SELECT COUNT(*)
   FROM Thread WHERE is_poll='y') as
   VoteActiveness
FROM Vote
```

All types of activeness can be calculated during the lifetime or different periods of time, for example by years.

Here is the query for calculating activeness by year:

```
SELECT
   member_id,
   Year(evdate) as Years,
   COUNT(*)*100.0/(SELECT COUNT(*)
   FROM Post) AS PostActivenessYear
FROM Post
GROUP BY member_id, Year(evdate)
```

Attractiveness is calculated in two ways. The first way: attractiveness is the arithmetic mean of all divisions' of members' reactions on the i-th member's posts divided by the total number of forum's members, divided by the number of i-th member's posts.

$$Attractiveness(member_i) = \frac{\sum_{j}^{N} \frac{card(Feedback(post_j(member_i)))}{card(Member)}}{card(Post(member_i))},$$

where $Feedback(post_j(member_i))$ – replies of all members to j-th post created by the i-th member;

$Member$ – a set of forum members;

$Post(member_i)$ – number of posts created by i-th member.

Query for the first way of calculation of attractiveness looks as follows:

```
SELECT
   F.member_id,
   COUNT(*)*100.0/(SELECT COUNT(*)
   FROM Member)/(SELECT COUNT(*) FROM
   Post WHERE P.member_id=F.member_id)
   as Attractiveness
FROM Feedback as F
GROUP BY F.member_id
```

The second way of calculating the attractiveness is: as a relation of the number of replies in the threads, created by a member, to the total replies count in all threads.

$$Attractiveness(member_i) = \frac{card(Reply(Thread(member_i)))}{card(Reply(Thread))},$$

where $Reply(Thread(member_i))$ – replies in the threads, created by a i-th member

$Reply(Thread)$ – total replies count in all threads.

Query for the second way of calculation of attractiveness looks as follows:

```
SELECT
   member_id,
   SUM(replies)*100.0/(SELECT
   SUM(replies) FROM Thread) as
   Attractiveness
FROM Thread
GROUP BY member_id
```

Creativeness is calculated in two ways. The first way: Creativeness is the arithmetic mean of all divisions' positive feedbacks divided by the total number of feedbacks to each post created by the member.

$$Creativeness(member_i) = \frac{\sum_j^N \frac{card(Feedback_{positive}(post_j(member_i)))}{card(Feedback_{total}(post_j(member_i)))}}{card(Post(member_i))},$$

where $Feedback_{positive}(post_j(member_i))$ – all positive feedbacks to the j-th post made by i-th member;

$Feedback_{total}(post_j(member_i))$ – all feedbacks to the j-th post made by i-th member;

$Post(member_i)$ – all posts created by i-th member.

Query for the first way of calculation of creativeness looks as follows:

```
SELECT
   F1.member_id,
   COUNT(*)*100.0/(SELECT COUNT(*)
   FROM Feedback as F2
   WHERE F2.member_id=F1.member_id)/
   (SELECT COUNT(*) FROM Post as P
    WHERE P.member_id=F1.member_id) as
   Creativeness
FROM Feedback as F1
WHERE F1.value=1
GROUP BY F1.member_id
```

The second way of calculating the member creativeness is: as a relation of the sum of views of all threads, created by a member, divided by the total number of all views of all forum threads.

$$Creativeness(member_i) = \frac{card(View(Thread(member_i)))}{card(View(Thread))},$$

where $View(Thread(member_i))$ – number of views of threads, created by the i-th member,

$View(Thread)$ – total number of views of all threads.

Query for the second way of calculation of creativeness looks as follows:

```
SELECT
  member_id,
  SUM(views)*100.0/(SELECT SUM(views)
  FROM Thread) as Creativeness
FROM Thread
GROUP BY member_id
```

Results of all queries are stored in the table *member_feature* that stores value characteristics of members (Fig. 4).

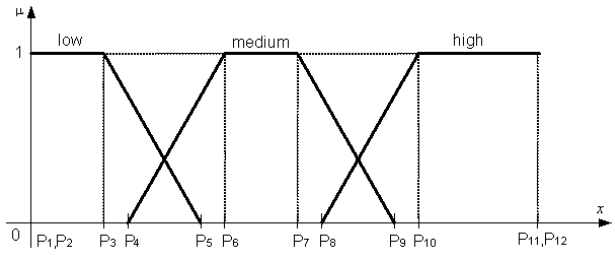


Fig. 4 – Linguistic variable

3.5. CALCULATION OF VALUES OF LINGUISTIC VARIABLES AND BELONGING MEASURES

Further we build linguistic variables based on the corresponding fuzzy variables in such a way.

Linguistic variable is given as a tuple $\langle \beta, T, X, M \rangle$, where β – name of variable, T – domain of variable – set of names of fuzzy variables. $T = \{"low","medium","high"\}$, X – domain of fuzzy variables, that describe linguistic variable; $X = [0;1]$; M – set of membership measures of fuzzy variables.

The first step in calculation of values of linguistic variables is defining the interval of linguistic variable to which the member's characteristic belongs.

In the information system, for defining this interval *BorderDefinition* (*feature*, *numeric_value*) function was realized. This function is based on the following query.

```
SELECT
  p1.feature_variable_id,
  p1.value as leftvalue,
  t1.value as lefttype,
  p2.value as rightvalue,
  t2.value as righttype
FROM feature_parameter as p1
INNER JOIN feature_variable as v on
  v.id = p1.feature_variable_id
INNER JOIN parameter_type as t1 on
  p1.parameter_type_id = t1.id
INNER JOIN parameter_type as t2 on
  t2.value = t1.value+1
INNER JOIN parameter as p2 on
  p2.parameter_type_id = t2.id AND
  p2.feature_variable_id =
  p1.feature_variable_id
WHERE
```

```
  v.feature_id = feature AND
  p1.value = (SELECT MAX(p2.value)
  FROM feature_parameter as p2
  WHERE p2.feature_variable_id =
  p1.feature_variable_id AND p2.value
  <= numeric_value)
```

This is an assisting function that helps to define fuzzy values of linguistic variables. For defining of fuzzy values of linguistic variables authors created a function *LinguisticName*(*feature*, *numeric_value*), which is based on these two queries:

```
INSERT INTO Measures(
  feature_variable_id,
  belonging_measure)
SELECT
  feature_variable_id,
  CASE
    WHEN lefttype=1 AND
    rightvalue<>leftvalue
    THEN (numeric_value -
    leftvalue)/(rightvalue -
    leftvalue)
    WHEN lefttype=2 OR
    rightvalue=leftvalue
    THEN 1
    WHEN lefttype=3 AND
    rightvalue<>leftvalue
    THEN (rightvalue -
    numeric_value)/(rightvalue -
    leftvalue)
  END as belonging_measure
FROM BorderDefinition(feature,
numeric_value)

SELECT TOP 1
  feature_variable_id
FROM Measures as m
INNER JOIN feature_variable as v on
  v.id = m.feature_variable_id
WHERE belonging_measure = (
  SELECT MAX(belonging_measure)
  FROM Measures)
ORDER BY ranking DESC
```

Next step is figuring out the belonging measure of the member to the value of the linguistic variable.

With this purpose in mind, the function *BelongingMeasure*(*feature*, *numeric_value*) was created and it is based on the following queries

```
INSERT INTO Measures(
  feature_variable_id,
  belonging_measure)
SELECT
  feature_variable_id,
  CASE
    WHEN lefttype=1 AND
    rightvalue<>leftvalue
    THEN (numeric_value -
    leftvalue)/(rightvalue -
    leftvalue)
    WHEN lefttype=2 OR
    rightvalue=leftvalue
    THEN 1
    WHEN lefttype=3 AND
    rightvalue<>leftvalue
    THEN (rightvalue -
    numeric_value)/(rightvalue -
    leftvalue)
  END as belonging_measure
FROM BorderDefinition(feature,
numeric_value)
```

and

```
SELECT MAX(belonging_measure) FROM
Measures
```

If a member's characteristic falls into the interval that is covered by two values of linguistic variable, then the member will get the value whose belonging measure is greater.

Linguistic variables and belonging measures values are calculated with the help of following query.

```
SELECT
  member_id,
  feature_id,
  LinguisticName(feature_id,
  numeric_value),
  BelongingMeasure (feature_id,
  numeric_value)
FROM member_feature
```

Query results are stored in the table *member_feature*.

Next step is classifying the members based on the classifying rules. Members' classification in the information system of Web-forums' members' behavior analyzing is realized on the basis of the rules that are stored in the data tables *class* and *rule* with the help of the following query

```
SELECT DISTINCT
  mf1.member_id,
  r1.class_id
FROM member_feature as mf1
INNER JOIN rule as r1 on
  r1.feature_variable_id =
  mf1.liguistic_value
WHERE r1.class_id NOT IN (
  SELECT r2.class_id
  FROM rule as r2
  LEFT JOIN member_feature as mf2 on
    mf2.linguistic_value =
    r2.feature_varible_id AND
    mf2.member_id = mf1.member_id
  WHERE mf2.member_id IS NULL)
```

Query results are stored in the table *member*.

Classification of members will be according to the following rules:

1. If Activeness(Member)= "high" and Creativeness(Member)= "high", then Member – Activist;
2. If Activeness(Member)= "high" and Loyalty(Member)= "high", then Member – Moderator;
3. If Activeness(Member)= "low" and Reactivity(Member)= "low" and

Creativeness(Member)= "high" and Attractiveness (Member)= "high", than Member – Author;

4. If Creativeness (Member)= "low" and Loyalty(Member)= "low" and Reactivity(Member)="high", then Member – Critic;
5. If Activeness(Member)= "high" and Creativeness(Member)= "low", then Member – Flamer.
6. If Activeness(Member)= "low" and Creativeness(Member)= "low", then Member – Reader.

4. CONCLUSION

During the information system development existing Web-forum CMS' were analyzed, process of Web-forums' members' behavior analyzing was researched and overviewed, information system database scheme was developed, and algorithms of Web-forum members behavior analyzing were programmed.

Development of information system of Web-forums' members' behavior analyzing and classifying is very important, since Web-forums are a very popular web-service and administrating them is a hard and complicated task.

A system developed by the means of DBMS Microsoft SQL Server. It allows Web-forum owners and administrators to analyze the behavior of Web-forums' members' and making the right decisions in the process of moderating a community.

5. REFERENCES

[1] Kravets. R. Typical ways of web-communities development / *Kravets. R., Peleschyshyn A.M., Syerov Yu. // Proceedings of the International Conference on Computer Science and Information Technologies, CSIT`2006, September 28th-30th*, Lviv, Ukraine, p.56–58.

[2] Syerov Yu.O. Simulation of behavior and classification of Web-communities participations of the base of fuzzy sets // *Transactions of National University "Lviv Polytechnics"*. – 2008. – #610 – Pp. 218-228. (in Ukrainian)

[3] Syerov Yu.O. *Fuzzy Sets Usage for Web-Community Participants Activity Simulation* // R.B. Kravets, Yu.O. Syerov / Transactions of Kharkiv national University of Radioelectronics "Automatic control system and automatic devices", issue. 141, 2007. – Pp. 113-118. (in Ukrainian)

[4] Kruglov V. V. *Fuzzy Logic and Artificial Neural Networks: Tutorial* / Kruglov V. V., Dli M. I., Golunov R. Yu. // – Moscow: Publishing House of Physical and Mathematical Literature, 2001. – 224 p. – ISBN 5-94052-027-8. (in Russian)

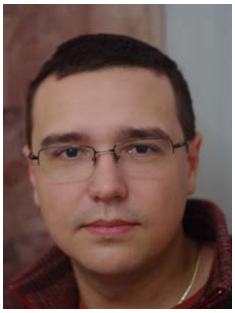
[5] Welser H. T. *Visualizing the Signatures of Social Roles in Online Discussion Groups* / Howard T. Welser, Eric Gleave, Danyel Fisher, Marc Smith // http://www.cmu.edu/joss/content/articles/volume8/Welser/

[6] Brandes U *Exploratory Network Visualization: Simultaneous Display of Actor Status and Connections* / Ulrik Brandes, Jörg Raab, Dorothea Wagner // [Online] http://www.cmu.edu/joss/content/articles/volume2/BrandesRaabWagner.html

[7] Freeman L. *Visualizing Social Networks*, / Linton C. Freeman, [Online] http://www.cmu.edu/joss/content/articles/volume1/Freeman.html

[8] Peleshchyshyn A.M. *Web 2.0 – the second chance for UANet [Electronic Resource]* Online Journal of T. Shevchenko Scientific Society, 2006. – [Online]: http://ntsh.org/uaweb2

[9] Pasichnyk V.V. *Web-community Efficiency Criteria Analysis on the Forums Base* / Pasichnyk V.V., Kravets R.B., Syerov Yu.O. // Procedings of 12th International Youth Forum "Radioelectronics and Youth in XXI century", 1-3 April 2008. Part 2, p. 493, Kharkiv. (in Ukrainian)



Yury Syerov, *master of computer sciences, National University "Lviv Polytechnics" 1998-2003.*

Assistant at department of Information systems and networks.

Scientific interests: Web-communities, organization of effective Web-communities, web mining technologies.

Ruslan Kravets, *Lviv Ivan Franko State University, 1990-1995.*

Associated Propfessor at department of Information systems and networks.

Scientific interests: intelligent systems, data mining technologies, text mining and web mining, databases and data warehouses.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

МЕТОД СТРУКТУРНОГО СИНТЕЗУ МЕРЕЖЕВИХ ПРИКЛАДНИХ ПРОЦЕСОРІВ

І. Майків [1)], А. Степаненко [1)], Д. Вобшал [2)]

1) Науково дослідний інститут Інтелектуальних комп'ютерних систем, Тернопільський національний економічний університет, 46009 Україна, м. Тернопіль, вул. Львівська 11, mim@tanet.edu.te.ua, {andrew.stepanenko@gmail.com}, http://www.ics.tneu.edu.ua/
2) Esensors Inc., 4240 Ridge Lea Road Suite 37 Amherst, NY 14226, USA 14226, darold@wobschall.com, http://www.eesensors.com/

Резюме: *на основі методу морфологічного аналізу та синтезу запропоновано метод структурного синтезу мережевих прикладних процесорів. Метод включає етапи функціонального аналізу, структурного синтезу, та пошуку множини оптимальних рішень і поєднує лексикографічний критерій переваги (L-критерій) для відбору електронних компонентів на етапі функціонального аналізу та безумовний критерій переваги (оптимальності по Парето, π-критерій) на етапі пошуку множини оптимальних рішень, що розглядаюся в літературі, як альтернативні методи пошуку оптимальних рішень. Поєднання L та π -критеріїв дозволяє зменшити число синтезованих альтернативних варіантів на етапі структурного синтезу, а всі отриманні рішення є допустимими.*

Ключові слова: *структурний синтез, морфологічний аналіз, лексикографічний критерій переваги, безумовний критерій переваги.*

ВСТУП

Сучасні вимірювально-керуючі системи (ВКС) реалізуються як багаторівневі системи із розподіленими обчислювальними ресурсами на базі промислових шин (інтерфейсів). Для їх уніфікації, національним інститутом стандартів США (*NIST*), прийнято серію стандартів IEEE-1451, що регламентує вимоги до апаратного та програмного забезпечення ВКС. Також, введено означення мережевого прикладного процесора (МПП) – вузла обробки даних, який займає проміжний рівень між сервером мережі (верхній рівень) і сенсорами та виконавчими механізмами (нижній рівень). Основними функціями МПП є: 1) поточна обробка даних; 2) підтримка набору послідовних інтерфейсів, по яких здійснюється обмін даними як із вузлами нижнього рівня, так і з сервером системи. Однак кожна область застосування характеризується своєю множиною функціональних задач (ФЗ), що вимагає застосування спеціалізованих МПП, технічні параметри яких оптимальні для даної області, а отже створення методу структурного синтезу МПП, оптимальних за функціонально-вартісними показниками.

1. АНАЛІЗ ВІДОМИХ МЕТОДИК СТРУКТУРНОГО СИНТЕЗУ

Задача оптимального структурного синтезу є однією із традиційних задач, загальні підходи до вирішення якої розглянуто в [1-7], більшість із яких базується на методах дискретної оптимізації (віток та границь, морфологічного синтезу) В [8] представлено метод структурного синтезу оптимальних апаратних структур багатопроцесорних інформаційних систем, який включає етапи аналізу задач, що вирішує система та саме структурного синтезу за результатами аналізу. Однак запропонований метод орієнтований на реалізацію систем на базі застарілих мікропроцесорів (МП) серії К580, К581 та ін. Виходячи із їх обмеженої продуктивності, його цільова функція полягає в визначенні необхідного числа МП та оптимальному розподілі між ними множини ФЗ, що реалізує система.

Однак сучасні МП та мікроконтролери (МК) характеризуються високою продуктивність (швидкістю виконання операцій), що лежить в межах від одиниць до сотень мільйонів операцій за секунду (*MIPS*), і цільова функція нового методу оптимального структурного синтезу

повинна забезпечити вибір МК, продуктивність якого буде достатньою для вирішення множини ФЗ. Тому метод запропонований в [8], не може безпосередньо використовуватись в процесі структурного синтезу МПП і потребує вдосконалення.

Альтернативний підхід до створення компонентів дистрибутивних комп'ютерних систем, за функціонально-вартісними показниками запропоновано в [9], який базується на методі морфологічного аналізу та синтезу [10, 11]. В процесі проектування системи розглядається ряд альтернативних рішень, що в цілому дозволяє зробити вибір між: 1) необхідністю нової розробки; 2) адаптацією вже відомих рішень; 3) компонуванням існуючих програмно-апарратних вузлів. Однак даний метод не містить етапу попереднього аналізу ФЗ об'єкта, який проектується, що не дозволяє об'єктивно оцінити вимоги як до апаратних так і програмних ресурсів, необхідних для ефективного функціонування ВКС. Вибір варіантів вузлів здійснюється на основі заданих технічних обмежень (ТО), які задовольняють технічне завдання (ТЗ). Також в процесі пошуку множини оптимальних рішень виконується повний перебір всіх можливих комбінацій вузлів, що часто неприйнятно збільшує час пошуку оптимальних рішень.

В [12] запропоновано метод наскрізного проектування портативних приладів та спеціалізованих мікропроцесорних систем на базі МК. Однак процес проектування орієнтовано на МК лише одного типу і не проводиться оцінка ефективності багатоваріантної реалізації на базі МК як різної архітектури так і виробників.

Таким чином, як показали результати аналізу, на даний час відсутній метод оптимального структурного синтезу цифрових систем на базі МК, що характеризуються широким набором функціональних можливостей та значною різницею в ціні.

Метою роботи є створення методу оптимального структурного синтезу МПП на базі МК виходячи із набору (множини) їх функціональних показників, що можуть виступати як критерії оцінки якості.

2. ЗАГАЛЬНІ ПІДХОДИ ДО ПРОЦЕСУ СТРУКТУРНОГО СИНТЕЗУ МПП

Метод структурного синтезу МПП, оптимальних за функціонально-вартісними показниками повинен включати етапи: 1) функціонального аналізу; 2) структурного синтезу; 3) пошуку структур оптимальних за функціонально-вартісними показниками.

Узагальнений алгоритм запропонованого методу (рис.1) включає 8 етапів, що умовно розділені на дві групи: функціонального аналізу та структурного синтезу.

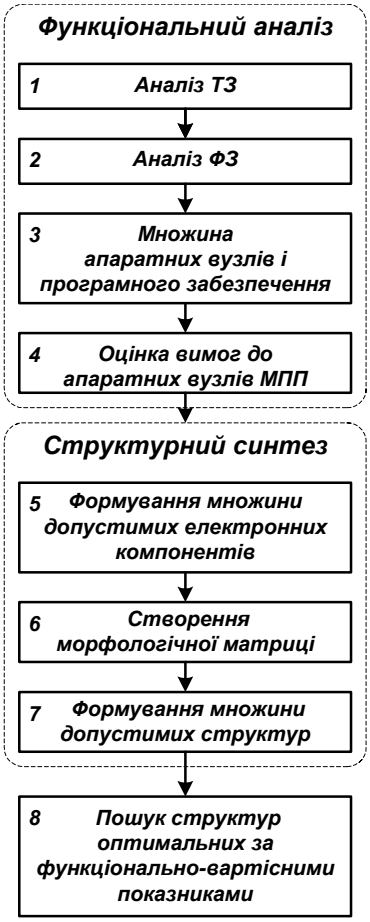


Рис. 1 – Алгоритм струкутрного синтезу мережевих прикладних процесорів

3. ФУНКЦІОНАЛЬНИЙ АНАЛІЗ МПП

На першому етапі процесу проектування виконується аналіз ТЗ на МПП, що дозволяє визначити перелік ФЗ, які він реалізує. В цілому ФЗ поділяють на системні та прикладні. Системні ФЗ реалізуються операційною системою або спеціалізованим програмним забезпеченням і забезпечують взаємодію прикладних ФЗ, визначають порядок їх виконання, розподіл пам'яті між ними і т.д. Прикладні ФЗ виконують обслуговування зовнішніх пристроїв та користувачів. Із загальної множини прикладних ФЗ наведеної в [8], МПП реалізує лише два класи: 1) передачі даних; 2) обробки даних.

Кількість прикладних ФЗ, що реалізує МПП, визначається множиною I, що уточнюється в процесі аналізу ТЗ на МПП

$$I = \{i : i = \overline{1, I}\}, \qquad (1)$$

де I – загальне число ФЗ, i – порядковий номер ФЗ.

На другому етапі виконується аналіз ФЗ. В результаті аналізу для кожної ФЗ визначається множина інформаційних параметрів, необхідних для оцінки об'єму вхідних/вихідних даних $\{l_i, \tau_i, n_i, t_i\}$, де: 1) l_i, – середній розмір вхідних повідомлень; 2) τ_i – періодичність рішення i-ї ФЗ; 3) n_i – середня кількість повідомлень, що поступає для i-ї ФЗ, за період τ_i; 4) t_i – час вирішення i-ї ФЗ. В результаті для множини I формується множина векторів (2), які задають набір функціональних обмежень для МПП

$$< l_i >, < \tau_i >, < n_i >, < t_i >, i = \overline{1, I}. \quad (2)$$

Ефективність реалізації МПП i-ї ФЗ визначається наступною умовою

$$t_i < \tau_i. \quad (3)$$

Будь-яка ФЗ є багатоетапним процесом перетворення даних, тому доцільно здійснити розподіл ФЗ на ряд окремих функціональних складових – функцій системи (ФС), кожна із яких реалізує один із етапів обробки даних, а взаємозв'язки між ними відображають процес перетворення даних. Перелік всіх ФС створює множину J

$$J = \{j : j = \overline{1, J}\}, \quad (4)$$

де J – загальне число типів ФС, j – порядковий номер ФС.

Взаємозв'язок між i-ю ФЗ та j-ю ФС представляється у вигляді матриці $[\xi_{ij}]$, $i = \overline{1, I}$ $j = \overline{1, J}$, в рядки якої записується перелік ФЗ, що реалізує МПП, а у стовпці – перелік всіх ФС. Кожен елемент матриці $[\xi_{ij}]$ є двійковим числом, що дорівнює 1, якщо для реалізації i-ї ФЗ необхідно ФС j-го типу, і 0 в протилежному випадку. Сукупність всіх ФС, необхідних для реалізації i-ї ФЗ, формує множину J_i, що належить множині J: $J_i \in J$. Також, через набір попередньо уточнених інформаційних параметрів для i-ї ФЗ $\{l_i, \tau_i, n_i, t_i\}$. визначаються параметри ФС: 1) середня довжина вхідних повідомлень – l_{ij} (біт); 2) інтенсивність вхідних повідомлень – λ_{ij} (повідомлення/сек.). В [8] наведено перелік основних типів ФС та аналітичні залежності для обчислення l_{ij} і λ_{ij}.

На третьому етапі визначається необхідне апаратне забезпечення (АЗ) та програмне забезпечення (ПЗ), що реалізує окремі ФС. Перелік всіх типів АЗ уточнюється в процесі аналізу ТЗ і створює множину H

$$H = \{h : h = \overline{1, H}\}, \quad (5)$$

де H - загальна кількість типів апаратних засобів.

Множину H розділяють за функціональним призначенням на ряд підмножини: 1) МК – множина H^{MC}; 2) пристроїв вводу/виводу – множина H^{IF}; 3) постійних запам'ятовуючих пристроїв (ПЗП) – H^{ROM}; 4) оперативних запам'ятовуючих пристроїв (ОЗП) – H^{RAM}. В цілому $H = H^{MC} \bigcup H^{IF} \bigcup H^{ROM} \bigcup H^{RAM}$.

В подальшому для j-ї ФС складається матриця відповідності АЗ $[\xi_{jh}]$, $j = \overline{1, J}$, $h = \overline{1, H}$, в рядки якої записується перелік ФС, що реалізує МПП, а у стовпці – перелік всіх типів АЗ. Кожний елемент матриці $[\xi_{jh}]$ є двійковим числом, що дорівнює 1, якщо для реалізації j-ї ФС необхідно використати апаратне забезпечення h-го типу, і 0 в протилежному випадку. Сукупність всіх апаратних засобів, необхідних для реалізації j-ї ФС, формує множину H_j, що належить множині H: $H_j \in H$.

Аналогічним чином формується перелік необхідного ПЗ, що створює множину S

$$S = \{s : s = \overline{1, S}\}, \quad (6)$$

де S - загальна кількість модулів ПЗ.

Також формується матриця відповідності ПЗ $[\xi_{js}]$, $j = \overline{1, J}$, $s = \overline{1, S}$, в рядки якої записується перелік ФС, що реалізує МПП, а у стовпці – перелік всіх модулів ПЗ. Кожен елемент матриці $[\xi_{js}]$ є двійковим числом, що дорівнює 1, якщо для j-ї ФС необхідно використати ПЗ s-го типу, і 0 в протилежному випадку. Сукупність всіх модулів ПЗ необхідних для реалізації j-ї ФС, формує множину S_j, що належить множині S: $S_j \in S$. Множина АЗ та ПЗ, необхідного для реалізації i-ї ФЗ, визначається наступними співвідношеннями

$$H_i = \sum_{j \in J_i} \sum_{h \in H_J} \xi_{ij} \cdot \xi_{jh}, \quad (7)$$

$$S_i = \sum_{j \in J_i} \sum_{s \in S_J} \xi_{ij} \cdot \xi_{js} \,. \qquad (8)$$

На четвертому етапі проводиться аналіз вимог до АЗ, зокрема: 1) продуктивності МК; 2) об'ємів ПЗП та ОЗП; 3) часових обмежень.

3.1 ОЦІНКА НЕОБХІДНОЇ ПРОДУКТИВНОСТІ МК

В роботі [13] показано, що МПП доцільно реалізувати як двох або багато процесорну структуру, в якій незалежно, на окремих МК, реалізуються процеси обробки та обміну даними. В цьому випадку, оцінка вимог до АЗ кожного вузла виконується окремо, по загальній методиці, враховуючи особливості їх функціонування.

Сукупність МК, що реалізують МПП, задається множиною $K = \{k : k = \overline{1, K}\}$, де K – кількість використаних МК.

Оцінку необхідної продуктивності МК проведемо, використавши математичний апарат теорії масового обслуговування (ТМО) [14]. В цьому випадку, кожен функціональний вузол можна розглядати як систему масового обслуговування (СМО) де МК виступає в якості обслуговуючого вузла, заявками на обслуговування є повідомлення, а процесом обслуговування заявки є процес обробки повідомлення (рис. 2).

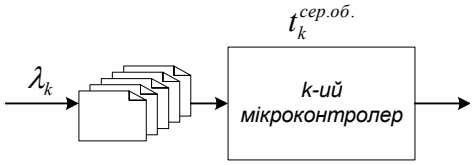


Рис. 2 – Модель СМО для оцінки продуктивності k-го мікроконтролера

Завданням є визначити продуктивність k-го МК $P_{\Sigma k}$, необхідну при максимальному потоці заявок на обслуговування. Оскільки, в більшості випадків, неможливо точно встановити максимальне значення потоку заявок, тому для спрощення оцінки використовують середню сумарну інтенсивність потоку заявок λ_k, що виконує k-й МК

$$\lambda_k = \sum_{i \in I_k} \sum_{j \in J_i} \lambda_{ij} \,. \qquad (9)$$

Тоді середній час обслуговування однієї заявки $t_k^{сер.об.}$ для k-го МК визначається як

$$t_k^{сер.об} = \frac{\sum_{i \in I_k} \sum_{j \in J_i} \lambda_{ij} \cdot l_{ij} \cdot \rho_i}{P_k \cdot \left(\sum_{i \in I_k} \sum_{j \in J_i} \lambda_{ij} \right)}, \qquad (10)$$

де I_k – множина ФЗ, що вирішує k-й МК; λ_{ij} – інтенсивність вхідних повідомлень (повідомлення/сек.); l_{ij} – середня довжина вхідних повідомлень (біт); ρ_j – питома продуктивність МК (операцій/біт), при виконанні j-ї ФС; P_k – реальна продуктивність k-го МК.

Коефіцієнт завантаження k-го МК визначається наступним співвідношенням

$$\gamma_k = \frac{P_{\Sigma k}}{P_k} = \lambda_k \cdot t_k^{сер.об.}, \qquad (11)$$

де $P_{\Sigma k}$ – необхідна продуктивність та P_k – реальна продуктивність k-го МК (оп./с).

Врахувавши λ_k та $t_k^{сер.об.}$ (9, 10), необхідна продуктивність для k-го МК визначається наступним співвідношенням

$$P_{\Sigma k} = \sum_{i \in I_k} \sum_{j \in J_i} \lambda_{ij} \cdot l_{ij} \cdot \rho_j \,, \qquad (12)$$

а на основі співвідношення (11) задається умова вибору типу k-го МК

$$P_k \geq \frac{P_{\Sigma k}}{\gamma_k} \,. \qquad (13)$$

Враховуючи, що для систем реального часу, згідно [8, 15] рекомендоване значення γ_k становить 0,5-0,6, тип k-го МК вибирають виходячи із умови, що

$$P_k = (1{,}6 \div 2) \cdot P_{\Sigma k} \,. \qquad (14)$$

3.2 ОЦІНКА НЕОБХІДНОГО ОБ'ЄМУ ПЗП

У ПЗП МК зберігається набір керуючих та прикладних програм. В процесі аналізу ФЗ та ФС, що реалізує k - й МК, визначається множина S_{Σ_k} необхідного ПЗ

$$S_{\Sigma_k} = S_{0_k} \cup S_{J_k} \,, \qquad (15)$$

де S_{0_k} – множина керуючих програм; S_{J_k} – множина прикладного ПЗ, що реалізує k-й МК.

Зазвичай МК реалізує одну керуючу програму, а прикладне ПЗ складається із

множини підпрограм, що забезпечують реалізацію k-тим МК множини ФС J_k. Тоді необхідний об'єм ПЗП Q_k для k-го МК визначається за наступним співвідношенням

$$Q_k = Q_{0_k} + \sum_{j \in J_k} Q_{j_k}, \qquad (16)$$

де Q_{0_k} – об'єм пам'яті необхідний для керуючої програми; $\sum_{j \in J_k} Q_{j_k}$ – об'єм пам'яті необхідний для зберігання прикладного ПЗ.

Однак, пропонований в [13] підхід до реалізації МПП забезпечує можливість його дистанційного перепрограмування, тому в керуючій програмі головного МК немає необхідності, а виконуване прикладне ПЗ може мати мінімально необхідну, в даний момент часу, множину підпрограм. Таким чином, пропонований в [13] МПП має суттєву перевагу в необхідному об'ємі ПЗП відносно відомих рішень.

Якщо ПЗ зберігається у внутрішньому ПЗП, тип МК, визначається співвідношенням

$$Q_k \le Q^{ROM}_{h \in H^{MC}}, \qquad (17)$$

де $Q^{ROM}_{h \in H^{MC}}$ - об'єм ПЗП МК h-го типу, що належить множині H^{MC}.

У випадку, якщо ПЗ зберігається у зовнішній пам'яті, необхідна кількість мікросхем N_k^{ROM} визначається співвідношенням

$$N_k^{ROM} = \left\rceil \frac{Q_k}{Q^{ROM}_{h \in \{H^{ROM} \cup H^{RAM}\}}} \right\lceil, \qquad (18)$$

де $Q^{ROM}_{h \in \{H^{ROM} \cup H^{RAM}\}}$ – об'єм пам'яті h-го типу, що належить множині, яка об'єднує підмножини ПЗП - H^{ROM} та ОЗП – H^{RAM}, що реалізують функцію ПЗП (див [13]); $\left\rceil \alpha \right\lceil$ – відношення α, заокруглене в більшу сторону до цілого числа.

3.3 ОЦІНКА НЕОБХІДНОГО ОБ'ЄМУ ОЗП

ОЗП використовується для тимчасового зберігання повідомлень, організації обмінників та стеків. Необхідний об'єм ОЗП V_k k-го МК визначається співвідношенням

$$V_k = V_k^{data} + V_k^{stack}, \qquad (19)$$

де V_k^{data} – об'єм ОЗП, необхідний для зберігання повідомлень і організації буферів обміну даними; V_k^{stak} – об'єм ОЗП, необхідний для організації стеків програм.

При виконанні k-тим МК множини J_k ФС всі повідомлення обробляються відповідним ПЗ або перебувають в черзі на опрацювання. Всю множину програм, що реалізує k-ий МК, можна розділити на дві групи: 1) обробки даних; 2) вводу/виводу даних. Для програм обробки даних, час обробки повідомлень і час їх перебування в черзі залежать лише від продуктивності вибраного МК. В той же час, при виконанні програм виводу даних, час обробки повідомлень та час очікування в черзі, а отже довжина черги, залежать не лише від продуктивності МК, складності програми, а також від швидкості виводу даних відповідним пристроєм виводу (швидкодії інтерфейсу). Враховуючи те, що продуктивність типових 8-ми розрядних МК лежить в межах 2÷25 MIPS, а в деяких досягає 100 MIPS, можна констатувати, що для програм виводу час обробки повідомлень зазвичай займає незначний відсоток від часу виводу повідомлень. Таким чином, довжина черги повідомлень, що виводяться з МК, в основному буде залежати від швидкодії пристрою виводу (інтерфейсу).

Використавши математичний апарат теорії масового обслуговування [14], виведемо залежності для оцінки об'єму пам'яті, необхідної для зберігання повідомлень, що перебувають в черзі на обробку та обробляються k-им МК, а також об'єму пам'яті необхідної для зберігання повідомлень, що перебувають в черзі на вивід.

Представимо k-тий МК і набір його пристроїв вводу/виводу, як дворівневу систему масового обслуговування (рис.3), в якій підсистема першого рівня (ліва частина рис.3) виконує обробку повідомлень, а підсистема другого рівня (права частина рис.3) виконує виведення повідомлень.

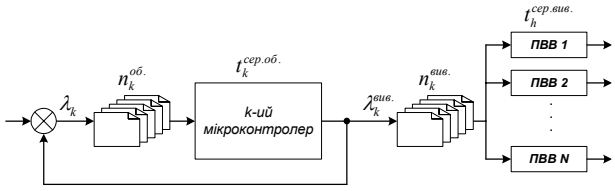


Рис. 3 – Модель СМО для оцінки об'єму ОЗП k-го мікроконтролера

В запропонованій моделі підсистема масового обслуговування першого рівня є одноканальною системою із необмеженою чергою на вході, де роль пристрою обслуговування заявок виконує k-тий МК. В ролі заявок виступають повідомлення, які поступають на обробку множиною програм S_{J_k}, що реалізують множину

J_k. Вхідний потік заявок λ_k є сумою вхідних повідомлень, що визначається за (9). Обслуговуванням заявки є процес обробки вхідного повідомлення k-тим МК, а середній час обслуговування однієї заявки $t_k^{сер.об.}$ визначається (10).

Підсистема другого рівня (див. рис. 3) є системою масового обслуговування, в якій кожен пристрій виводу (інтерфейс) можна розглядати як одноканальну систему із необмеженою чергою, яка працює на свого споживача. Вихідний потік $\lambda_k^{вив.}$ – сума повідомлень, що поступають від МК на вивід, визначається співвідношенням

$$\lambda_k^{вив.} = \sum_{i \in I_k} \sum_{j \in J_i} \lambda_{ij} \cdot \xi_{ij} \cdot \xi_{jh}, \qquad (20)$$

де ξ_{ij}, – двійкове число, що відображає зв'язок між i-ю ФЗ та j-ю ФС, а ξ_{jh} – двійкове число пов'язане із АЗ h-го типу, що належить множині H^{IF} і реалізує відповідний інтерфейс.

Означений потік повідомлень розділяється на множину потоків, число яких визначається кількістю пристроїв виводу (інтерфейсів) N_k^{IF}.

Об'єм пам'яті V_k^{data}, необхідний для зберігання даних, можна визначити за співвідношенням

$$V_k^{data} = l_k^{сер.об} \cdot n_k^{об.} + l_k^{сер.вив} \cdot n_k^{вив.}, \quad (21)$$

де $l_k^{сер.об}$, $l_k^{сер.вив.}$ – середня довжина повідомлень, що обробляються k-м МК та поступають на вивід; $n_k^{оч.}$ – середня кількість повідомлень, що перебувають в черзі на обслуговування; $n_k^{вив.}$ – середня кількість повідомлень, що перебувають в черзі на вивід.

Для підсистеми масового обслуговування, першого рівня, середня довжина повідомлень, що обробляються, $l_k^{сер.об}$ визначається співвідношенням

$$l_k^{сер.об} = \frac{\sum_{i \in I_k} \sum_{j \in J_i} \lambda_{ij} \cdot l_{ij}}{\sum_{i \in I_k} \sum_{j \in J_i} \lambda_{ij}}. \qquad (22)$$

Допускаючи безпріоритетний порядок обслуговування заявок, середнє число повідомлень, що перебувають в черзі на обслуговування k-тим МК, можна визначити за наступним співвідношенням

$$n_k^{об.} = \frac{\sum_{i \in I_k} \sum_{j \in J_i} t_{ijk} \cdot \gamma_k}{(1 - \gamma_k)}, \qquad (23)$$

де γ_k – коефіцієнт завантаження k-го МК; t_{ijk} – середній час обробки повідомлень j-ї ФС при виконанні i-тої ФЗ k-тим МК, що визначається наступним співвідношенням

$$t_{ijk} = \frac{l_{ij} \cdot \rho_j}{P_k}, \qquad (24)$$

де l_{ij} – середня довжина вхідних повідомлень; ρ_j – питома продуктивність МК при виконанні j-ї ФС; P_k – продуктивність k-го МК.

Для підсистеми масового обслуговування другого рівня середній час обслуговування заявок становить

$$t_k^{сер.вив.} = \frac{\sum_{i \in I_k} t_i^{сер.вив.} \cdot \lambda_i}{\sum_{i \in I_k} \lambda_i}, \qquad (25)$$

де λ_i – інтенсивність вихідних повідомлень; $t_i^{сер.вив.}$ – середній час виведення повідомлення по i-й ФЗ, що визначається співвідношенням

$$t_i^{сер.вив} = \frac{l_i}{\upsilon_{h \in H^{IF}}} \qquad (26)$$

де l_i – середня довжина вхідних повідомлень по i-й ФЗ; υ_h – пропускна швидкість інтерфейсу h-го типу, що належить множині H^{IF}.

Середня довжина повідомлень, що поступають на інтерфейс h-го типу при виконанні i-тої ФЗ визначається співвідношенням

$$l_i^{сер.вив.} = \sum_{j \in J_i} l_{ij} \cdot \xi_{jh}, \qquad (27)$$

де ξ_{jh} – двійкове число, що показує зв'язок між j-тою ФС, із апаратним забезпеченням інтерфейсу h-го типу, що належить множині H^{IF}.

Об'єм пам'яті V_k^{stack}, необхідний для організації стеків, є пропорційним кількості програм S_{J_k} і обчислюється за співвідношенням

$$V_k^{stack} = \sum_{j \in J_i} v_j^{stack}, \qquad (28)$$

де v_j^{stack} – об'єм пам'яті, необхідний для організації стека при виконанні ФС j-го виду.

Виходячи із попередніх результатів оцінки необхідного об'єму пам'яті, кількість мікросхем ОЗП N_k^{RAM} визначається співвідношенням

$$N_k^{RAM} = \left\rceil \frac{V_k^{data} - V_k^0}{Q_{h \in H^{RAM}}^{RAM}} \right\lceil, \qquad (29)$$

де V_k^0 – об'єм внутрішнього ОЗП МК; $Q_{h \in H^{RAM}}^{RAM}$ - об'єм ОЗП h-го типу, що належить множині, – H^{RAM}, $\rceil \alpha \lceil$ – відношення α, заокруглене в більшу сторону до цілого числа.

3.4 Оцінка часових обмежень

Одним із головних критеріїв оцінки ефективності пристроїв обробки даних є час рішення i-тої ФЗ, тобто проміжок часу, необхідний для перетворення вхідного повідомлення у вихідне, який можна представити як сумарний час виконання ФС, пов'язаних із даною ФЗ

$$t_i = \sum_{j \in J_i} t_{ij} \cdot \xi_{ij}. \qquad (30)$$

Зазвичай на значення t_i накладається ряд обмежень, пов'язаних із характером ФЗ, які вирішує МПП, що формує вектор обмежень $< t_i >, i = \overline{1, I}$.

При виконанні j-ої ФС використовується множина АЗ H_j, що дозволяє представити t_{ij} як сумарний час, затрачений АЗ на виконання необхідних функцій

$$t_{ij} = \sum_{h \in H_J} t_{ijh} \cdot \xi_{jh}, \qquad (31)$$

де t_{ijh} – час виконання АЗ h-го типу j-ої ФС, яка відноситься до i-тої ФЗ; ξ_{jh} – двійкове число, яке показує зв'язок між АЗ h-го типу та j-ою ФС.

В свою чергу, процес виконання ФС АЗ розділяють на етапи обробки та очікування, які відповідно потребують часу на обробку повідомлення $t_{ijh}^{об.}$, та часу очікування в черзі на обробку $t_{ijh}^{оч.}$. Тоді час виконання ФС в цілому визначається співвідношенням

$$t_{ijh} = t_{ijh}^{об.} + t_{ijh}^{оч.}. \qquad (32)$$

Вказані параметри $t_{ijh}^{об.}$ і $t_{ijh}^{оч.}$ визначаються параметрами АЗ. Для МК час обробки повідомлення визначається співвідношенням (24), тобто $t_{ijh}^{об.} = t_{ijk}^{об.}$. Час очікування повідомлень на обробку $t_{ijh}^{оч.}$ визначається середнім часом очікування для k-го МК за співвідношенням

$$t_{ijh}^{оч.} = t_k^{оч.} = \frac{n_k^{об.}}{\lambda_k}, \qquad (33)$$

де $n_k^{об.}$ – середня кількість повідомлень що перебувають в черзі на обслуговування k-тим МК (23); λ_k – середня сумарна інтенсивність потоку заявок на обробку k-тим МК (7).

Для пристроїв виводу даних (інтерфейсів) час обробки повідомлення визначається співвідношенням (26), тобто $t_{ijh \in H^{IF}}^{об.} = t_i^{сер.вив.}$.

4. СТРУКУТУРНИЙ СИНТЕЗ МПП

В результаті проведено функціонального аналізу визначено перелік необхідних АЗ (5) і структура ПЗ (6) МПП.

Також сформовано множину вимог до: 1) продуктивності МК (13, 14); 2) об'єму ПЗП (16, 18) та ОЗП (19, 21, 28, 29); 3) вектора часових обмежень для (25, 26), (30-33), що формує вектор технічних обмежень для МПП.

Це дозволяє перейти до наступного етапу проектування МПП – структурного синтезу, що передбачає формування множини допустимих варіантів структури МПП Ω_o та пошуку серед них множини оптимальних рішень Ω_o по заданому критерію якості.

Враховуючи те, що, множина доступних комплектуючих є достатньо великою на п'ятому етапі (див. рис.1) виконується процедура попереднього вибору електронних компонентів (ЕК), параметри яких відповідають множині ТО. Для цього доцільно використати один із формальних методів багатокритеріального відбору. Будь який ЕК характеризується множиною технічних параметрів, які формують множину показників якості

$$K_{як.} = \{k_{як.} : k_{як} = \overline{1, K_{як}}\} \qquad (34)$$

де $K_{як}$ – загальна кількість показників якості, $k_{як.}$-й показник якості.

Множину $K_{як}$ можна розділити на підмножини, що визначаються набором

технічних параметрів ЕК: 1) електричні – $K_{як.}^{ел.}$; 2) динамічні – $K_{як.}^{дин.}$; 3) конструктивні – $K_{як.}^{кон.}$; 4) експлуатаційні – $K_{як.}^{екс.}$

$$K_{як.} = K_{як.}^{ел} \bigcup K_{як}^{дин.} \bigcup K_{як.}^{кон.} \bigcup K_{як.}^{екс.}. \quad (35)$$

Виходячи із того, що вибір ЕК здійснюють за множиною технічних параметрів, які повинні відповідати множині ТО, доцільно для кожної групи ЕК сформувати відповідну множину параметрів $K_{від.}$, за якою здійснюється відбір ЕК

$$K_{від.} = \{k_{від.}, k_{від.} = \overline{1,G}\}, \quad (36)$$

де G – загальна кількість параметрі, за якими здійснюється відбір ЕК, $k_{від.}$-й показник якості, за яким здійснюється відбір ЕК.

Множини допустимих ЕК формуються шляхом використання лексикографічного критерію переваги (L-критерію) [2, 16], який є умовним критерієм, де в якості додаткової інформації особа, що приймає рішення (ОПР), вводить порядок пріоритету показників якості. Якщо показники якості співвідносяться між собою відношенням переваги ($\succ$), то введений ОПР пріоритет показників якості $k_{від.1} \succ k_{від.2} \succ ... \succ k_{від.G}$ задає строгий лінійний порядок $< k_{від.1}, k_{від.2}, ..., k_{від.G} >$, що визначає послідовність процедури вибору. В результаті векторна задача оптимального вибору за сукупністю показників якості перетворюється в скалярну, шляхом її розподілу на ряд оптимізаційних задач, що виконуються послідовно. Це дозволяє створити підсистему попереднього (дооптимізаційного) усунення ЕК, які характеризуються: 1) недостатнім рівнем технічних параметрів, і не задовольняють ТО на проектований МПП; 2) явно надлишковими технічними параметрами. Такий попередній відбір дозволить майже на порядок скоротити кількість оцінюваних при оптимізації варіантів МПП.

Оскільки для реалізації МПП використовують набір типових ЕК, що формують множину H (5), то використання L-критерію $L < k_{від.1}, k_{від.2}, ..., k_{від.G} >$ дозволяє сформувати множину кращих альтернатив Ω_L із множини можливих варіантів ЕК для кожної групи АЗ, $\Omega_L = \Omega_L^1 \bigcup \Omega_L^2 \bigcup ... \bigcup \Omega_L^H$, де H – загальна кількість типів ЕК АЗ, необхідних для реалізації МПП.

Відбір базується на послідовному виконанні умов (37) на множині можливих альтернативних варіантів ЕК h-го типу, що утворюють множину $\Omega^h = \{\omega_i^h : i = \overline{1, N^h}\}$, де ω_i^h - i-й варіант ЕК h-го типу, N^h – загальна кількість ЕК h-го типу.

$$\begin{cases} k_{від.1}(\omega_i^h) < k_{від.1}(\omega_j^h) / \omega_i^h \succ \omega_j^h \\ k_{від.1}(\omega_i^h) = k_{від.1}(\omega_j^h); k_{від.2}(\omega_i^h) < k_{від.2}(\omega_j^h) / \omega_i^h \succ \omega_j^h \\ ..... \\ k_{від.t}(\omega_i^h) = k_{від.t}(\omega_j^h), t = \overline{1,(G-1)}; k_{від.G}(\omega_i^h) < k_{від.G}(\omega_j^h) / \omega_i^h \succ \omega_j^h \end{cases} \quad (37)$$

Оскільки МПП є достатньо складною технічною системою, яка характеризується множиною технічних параметрів та обмежень, доцільно використати одну із відомих методик багатокритеріальної векторної оптимізації [2, 6], для пошуку множини оптимальних рішень.

Для її реалізації на шостому етапі, запропонованого методу, проводиться декомпозиція МПП із використанням морфологічного методу [10, 11, 17], який передбачає виділення найважливіших ФС, що реалізуються множиною базових компонентів

$$N = \{n; n = \overline{1, N}\}, \quad (38)$$

де N – число базових компонентів з яких формується морфологічна матриця (ММ), із кількістю рядків N (число базових вузлів) та кількістю стовпців M (максимальна кількість альтернативних варіантів реалізації вузлів)

$$\begin{Bmatrix} P_1 =< P_{11} \dots P_{1j} \dots P_{1M} > \\ \vdots \\ P_i =< P_{i1} \dots P_{ij} \dots P_{iM} > \\ \vdots \\ P_N =< P_{N1} \dots P_{Nj} \dots P_{NM} > \end{Bmatrix} \quad (39)$$

де $i = 1 \dots N$, $j = 1 \dots M$.

Наступні етапи – формування множини допустимих структур та пошуку множини структур МПП оптимальних за функціонально-вартісними показниками, базуються на основних положеннях теорії дискретної оптимізації [18], якими передбачається, що кожний базовий

варіант вузла, множина яких утворена після декомпозиції МПП, характеризується множиною показників якості (36), що формують вектор $\overrightarrow{K_{від.}} =< k_{від.1}, k_{від.2,} \ldots, k_{від.G} >$.

Основними критеріями оптимізації є: 1) собівартість вузла C_{nm} – визначається затратами на реалізацію m-го варіанту n-го вузла; 2) показник функціональної ефективності вузла E_{nm}. Для m-го варіанту n-го вузла значення показника E_{nm} визначається співвідношенням:

$$E_{nm} = \alpha_n \cdot (100 - E_m) \qquad (40)$$

де α_n – коефіцієнт, що відображає вагу вузла в МПП; E_m – відносна ефективність вузла у відсотках.

Враховуючи, що на попередніх етапах проведено відбір ЕК, технічні параметри яких відповідають множині обмежень заданих ТЗ та отриманих на етапі функціонального аналізу, то для оптимізації структури МПП достатньо двох критерії оцінки ефективності m-го варіанту n-го вузла (C_{nm}, E_{nm}). В цілому для МПП можна записати

$$C_\Sigma = \sum_{n=1}^{N}\sum_{m=1}^{M} C_{nm} \cdot \xi_{nm}, \qquad (41)$$

$$E_\Sigma = \sum_{n=1}^{N}\sum_{m=1}^{M} E_{nm} \cdot \xi_{nm}, \qquad (42)$$

де ξ_{nm} – двійкове число, що дорівнює 1, при виборі m-го варіанту n-го вузла.

Умовами оптимізації є: 1) альтернативність рішень за сторонами; 2) перевага одних рішень над іншими.

Перша умова (альтернативність) передбачає, що всі M можливих варіантів реалізації вузла в межах n-го рядка ММ є альтернативним

$$\sum_{m=1}^{M} \xi_{nm}, \text{ для } n = 1 \ldots N. \qquad (43)$$

Друга умова (перевага одних рішень над іншими) представляється наступним чином: на скінченій множині $P = P_1 \bigcup P_2 \ldots \bigcup P_N$, що задана ММ (39). Необхідно знайти вектор

$$\overrightarrow{X} =< j_1 \ldots j_i \ldots j_N > \in \{X\}. \qquad (44)$$

де $\{X\}$ – скінчена множина, сила якого визначається множиною всіх можливих комбінацій індексів j_i.

Вектор $\overrightarrow{X}$ визначає конфігурацію вибірки (одну із можливих комбінацій індексів j_i), яка забезпечує мінімізацію цільової функції, за критеріями (41) і (42), що в загальному виражається наступним співвідношенням:

$$F_{opt} = f(C_\Sigma, E_\Sigma) \to \min. \qquad (45)$$

Усі відомі методи двокритеріальної оптимізації зводять векторний синтез до скалярного [2, 6, 7]. Найбільш поширене [6] знаходження оптимального рішення за:

1) результуючим показником якості k_p як мінімумом добутку критеріїв якості:

$$k_p = C * E \to \min. \qquad (46)$$

або зваженої їх суми (коефіцієнти c_1 та c_2 задаються експертом):

$$k_p = (c_1 * C + c_2 * E) \to \min. \qquad (47)$$

2) показником ефективності (розглядається як функція собівартості та функціональної ефективності), при цьому пряма задача полягає в знаходженні системи з $E_{\min}$ при $C \le C_{\max}$, а обернена задача – у знаходженні $C_{\min}$ при заданому обмеженні $E \le E_{\max}$;

3) мінімаксним методом, при якому критерії синтезованих варіантів нормуються:

$$C_j = \frac{C_j - C_{\min}}{C_{\max} - C_{\min}}, \qquad E_j = \frac{E_j - E_{\min}}{E_{\max} - E_{\min}}. \qquad (48)$$

де, для кожного j варіанту вибирається більше значення з нормованих критеріїв $R_j = \max(C_j, E_j)$, а далі вибирається система з найменшим значенням $R = \min(R_j)$.

Протиріччя "вартість-функціональні можливості" виключає рішення, краще за обома критеріями, тому доцільне поєднання знань, досвіду та інтуїції експерта-розробника з автоматизованим синтезом множини не гірших рішень (МНР). При цьому спочатку згідно безумовного критерію переваги (БКП) (принцип оптимальності за Парето) [2] знаходять множину не гірших рішень (МНР), а потім з отриманої МНР експерт вибирає кінцеве рішення з допомогою умовного критерію переваги (46-48) з врахуванням ТО і перспективи використання МПП. Згідно із БКП приймається, що, варіант ω_i домінує над варіантом ω_j, якщо для кожного з критеріїв якості $k_l \in K_{від.}$ виконується умова:

$$k_l(\omega_i) \le k_l(\omega_j) \qquad (49)$$

5. ВИСНОВКИ

На основі методу морфологічного аналізу та синтезу представлено метод структурного синтезу мережевих прикладних процесорів, що включає етапи функціонального аналізу, структурного синтезу, та пошуку множини оптимальних рішень. Запропонований метод поєднує лексикографічний критерій переваги (L-критерій) для відбору електронних компонентів на етапі функціонального аналізу та безумовний критерій переваги (оптимальності по Парето, π-критерій) на етапі пошуку множини оптимальних рішень, що розглядаюся в літературі як альтернативні методи пошуку оптимальних рішень. Поєднання L та π – критеріїв дозволяє зменшити число синтезованих альтернативних варіантів на етапі структурного синтезу.

Отримане формалізоване рішення задачі дискретної оптимізації МПП, що є універсальним для широкого кола задач оптимального структурного синтезу цифрових пристроїв.

ПОДЯКА

Дана робота виконана за підтримки фонду цивільних досліджень і розвитку США, при виконання міжнародного українсько-американського проекту #UKC2-005073-KV-07, "Dynamically reprogrammable network control application processor with Internet capability".

6. СПИСОК ЛІТЕРАТУРИ

[1] Норенков И.П. *Основы автоматизированного проектирования*: *Учеб. для вузов. 2-е узд*. – М.: Изд-во МГТУ им. Н.Є. Баумана, 2002. – 366 с.

[2] Кандырин Ю.В. *Методы и модели многокритериального выбора вариантов в САПР: Учеб. Пособие для вузов*. – М.: Изд-во МЭИ, 2004. – 172 с.

[3] Половинкин А.И. *Основы инженерного творчества: Учеб. пособие для студентов втузов.* – М.: Машиностроение, 1988. – 368 с.

[4] Подиновский В.В., Ногин В.Д. *Парето-оптимальное решение многокритериальных задач*. – М.: Наука, 1982. – 256 с.

[5] Сергиенко И.В., Капшицкая М.Ф. *Модели и методы решения на ЭВМ комбинаторных задач оптимизации*. – К.: Наукова думка, 1981. – 288 с.

[6] Гуткин Л.С. *Оптимизация радиоэлектронных устройств по совокупности показателей качества*. – М.: Сов. радио, 1975. – 368 с.

[7] Саченко А.О. *Теория дискретной оптимизации компонентов корректирующего канала ИИС для измерения температуры* // Н.-т. сборник "Метрологическое обеспечение производства и КИР", 1988, С. 95-101.

[8] *Выбор микроЭВМ для информационных систем: Учеб. Пособие для втузов* / Н.М. Соломатин, Р.П. Шертвитис, М.М. Макшанцев. – М.: Высш. шк., 1987. – 120 с.

[9] Кочан В.В., Тимчишин В.О. *Синтез оптимальних структур низьковартісних комп'ютерних систем* // Вісник ДУ "Львівська політехніка". – 1998. – № 356. – С. 134 -144.

[10] Одрин В.М. *Методы морфологического анализа технических систем*. – М.: ВНИИПИ, 1989. – 312 с.

[11] *Морфологический синтез систем: морфологические методы поиска* / Одрин В.М. – Киев, 1986. – 40с. – (Перепринт / АН УССР. Ин-т кибернетики им. В.М. Глушкова; 86-5).

[12] Палагін О.В., Романов В.О., Галелюка І.Б. *Віртуальні методи проектування складних систем: оцінка надійності* // Інформаційні технології та комп'ютерна інженерія. – 2005. – № 3. – С. 147–150.

[13] I. Maykiv, A. Stepanenko, D. Wobshal, V. Kochan, R. Kochan, A. Sachenko Remote Reprogrammable NCAPs: Issues and Approaches // Proceedings of 4th IEEE international workshop on Intelligent DataAcquisition and Advancing ComputingSystems, 6-8 September, 2007, Dortmund, Germany, pp.109-113.

[14] Клейнрок Л. Теория массово обслуживания. – М.: Машиностроение, 1979. – 408 с.

[15] Phillip A. Laplante Real-Tome Systems Design and Analysis. Third edition/ – IEEE Press, Wiley-Interscince, 2004. – p. 506.

[16] Кандырин Ю.В. *Автоматизированный многокритериальный выбор альтернатив в инженерном проектировании*. – М.: Изд-во МЭИ, 1992. – 52 с.

[17] Одрин В.М., Картавов С.С. Морфологический анализ систем. Построение морфологических таблиц. – К.: Наукова думка, 1977. – 148 с.

[18] Колдасов Г.Д. *Оптимизация решений при морфологическом синтезе технических систем.* / Механизация и автоматизация управления, 1987, № 3, С. 3 – 6.

Ігор Майків *закінчив державний університет Львівська політехніка, по спеціальності "Радіотехніка" і отримав диплом радіоінженера (1996). Працював інженером-конструктором відділу спецтехніки на Тернопільському радіозаводі "Оріон" (1998-2002). Навчання в аспірантурі за спеціальністю 05.13.05 – Комп'ютерні системи та компоненти (2005-2009). В даний час працює молодший науковий співробітник НДІ Інтелектуальних Комп'ютерних систем Тернопільського національного економічного університету (ТНЕУ).*

Наукові інтереси: дослідження і створення цифрових систем на базі МК та програмованих логічних матриць.

Андрій Степаненко *закінчив у 2005 році магістратуру ТНЕУ за спеціальністю Комп'ютерні системи та мережі. Продовжив своє навчання в аспірантурі за спеціальністю 05.13.06 "Інформаційні технології".*

Наукові інтереси: методи і засоби автоматизованої верифікації програмного забезпечення інтелектуальних сенсорних мереж, розробка спеціалізованих мов програмування для вбудованих апаратних платформ, інженерія програмного забезпечення загалом

Дарольд Вобшал*, президент компанії Esensors Inc; м. Амхерст, шт. Нью-Йорк, США. Отримав ступінь доктора наук по біофізиці від Університету шт. Нью-Йорк в Баффало. Має більше 50-ти років досвіду розробки і побудови сенсорів та сенсорних систем. Виступає співзасновником та одним з головних експертів групи, яка розробляє стандарт IEEE 1451.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

A METHOD FOR STRUCTURAL SYNTHESIS OF NETWORK CAPABLE APPLICATION PROCESSORS

I. Maykiv [1)], A. Stepanenko [1)], D. Woschall [2)]

[1)] Research Institute of Intelligent Computer Systems, Ternopil National Economic University,
46009 Ukraine, Ternopil, 11 Lvivska str.,
mim@tanet.edu.te.ua, {andrew.stepanenko@gmail.com}, http://www.ics.tneu.edu.ua/
[2)] Esensors Inc., 4240 Ridge Lea Road Suite 37 Amherst, NY 14226, USA 14226,
darold@wobschall.com, http://www.eesensors.com/

Abstract: *A method for structural synthesis of Network Capable Application Processors (NCAPs) is proposed. It is based on a method of morphological analysis and synthesis and includes phases of functional analysis, structural synthesis, and search for a set of optimal solutions. The proposed method combines lexicographical criterion of preference (L- criterion) at a stage of functional analysis and unconditional criterion of preference (Pareto optimality) during the search phase, which are considered in literature as alternative methods of search for optimal solutions. Combining of criteria, makes it possible to reduce the number of synthesized alternative variants at a stage of structural synthesis and all obtained solutions are allowable.*

Keywords: *structural synthesis, morphological analysis, lexicographical criterion of preference, unconditional criterion of preference.*

Modern measurement and control systems are implemented as multilevel systems with the distributed computing resources on the basis of industrial buses. The IEEE-1451 series of standards has been developed by IEEE for unification of requirements for hardware and software of such systems. Also a term *Network Capable Application Processor (NCAP)* as a data processing device is defined by this set of standards. NCAP occupies an intermediate level between a network server (high level) and sensors and actuators (low level). The main functions of the NCAP are data processing and support of a set of serial interfaces. However, each area of application is characterized by the set of functional tasks which demand use of the specialized device technical parameters, which are optimized for the specified area of use.

The analysis of available literature on this subject has shown, that at present the methods for optimum structural synthesis of digital systems based on microcontrollers which are characterized by a wide set of features with a significant price range is absent. The objective of this work is creation of a method for optimum structural synthesis of NCAP on the basis of microcontrollers, with optimal cost-functionality relation obtained from a set functional parameters which can appear as criterion for estimation of the device quality.

The effective technique of structural synthesis of devices with optimal cost-functionality relation includes the next stages: 1) functional analysis; 2) structural synthesis; 3) search for optimal structures. The generalized algorithm of the proposed technique includes eight stages (Fig.1) which are divided into groups: functional analysis and structural synthesis.

The requirement specification for the NCAP is analyzed at the first stage of the design process. At this stage the list of functional tasks which are executed by the NCAP is created. Functional tasks are divided on system and application tasks. System tasks are realized by operational system (OS) or the specialized software and provide interaction for the application tasks. Application tasks serve external devices and users and are divided into two groups: 1) data exchange; 2) data processing.

The analysis of application tasks is performed at the second stage. The set of information parameters necessary to make estimation on the amount of input and output data for each application task is defined as a result of the analysis. As a result of this stage a set of vectors specifying functional restrictions for the NCAP is created.

The hardware and software necessary for implementation of all application functions of system is determined at the third stage as result of the requirement specification analysis and analysis of the list of application functions executed by the NCAP.

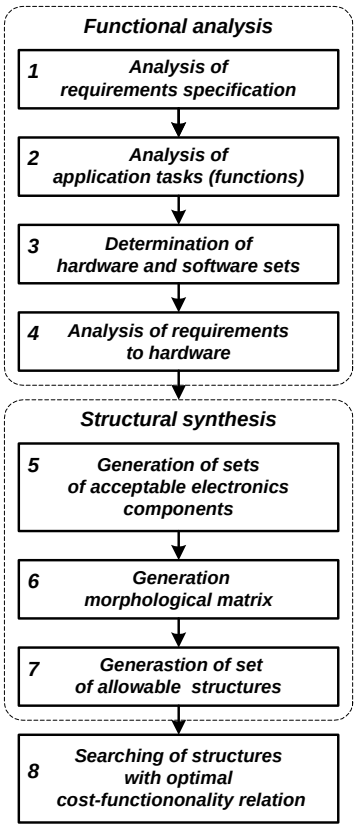


Fig.1 – Algorithm of method of structural synthesis

At the fourth stage the analysis of requirements for the device hardware is performed in particular: 1) processing power of a microcontroller; 2) memory size of ROM and RAM; 3) timing performances and temporal constraints. The estimation is made using elements of the theory of queues where the microcontroller is considered as the processing node, messages considered as service requests and message processing corresponds to service requests processing.

At the fifth stage the selection of electronic components is performed on the conducted functional analysis, where the list of the required hardware and software was created with timing characteristics and restrictions as well. The set of possible electronic components is created using lexicographic criterion of preference, which is a conditional criterion. A decision making person enters quality indicators priority list as additional information. As a result vector task of optimal selection of electronic components is transformed into a scalar task. This allows creation of a subsystem of a preliminary selection of electronic components which: 1) do not meet technical restrictions for the device being designed; 2) posses considerably better technical parameters. Such a preliminary selection enables to decrease the number of selection choices almost to the order of magnitude.

At the sixth stage a decomposition of the device into a set of basic nodes is performed using morphological method. Then a morphological matrix is created (NxM), where N is the number of basic nodes, and M is the maximum number of alternate variants of the N-th node.

At the seventh and eighth stages the set of possible device structures is generated among which the search for optimal cost-functionality combination is performed. The process of this set creation is based on principals of discrete optimization, where each node variant is characterized by implementation costs C_{nm} and functional effectiveness E_{nm}. In general for the device the following is true

$$C_\Sigma = \sum_{n=1}^{N}\sum_{m=1}^{M} C_{nm}\cdot\xi_{nm}\ ,\ E_\Sigma = \sum_{n=1}^{N}\sum_{m=1}^{M} E_{nm}\cdot\xi_{nm}$$

In this case the criterion function is described by the following equation $F_{opt} = f(C_\Sigma, E_\Sigma) \to \min$.

The known methods of bi-criterion optimization transform vector synthesis into scalar synthesis. Widespread is the search for optimal solution by:
1) resulting quality indicator as a multiply of quality criteria $k_p = C * E \to \min$ or weighted sum $k_p = (c_1 \cdot C + c_2 \cdot E) \to \min$; 2) effectiveness indicator (is considered as a function of self-cost and functional effectiveness). In this case the direct task lies in selection of a device with $E_{\min}$ and set restriction $C \le C_{\max}$, and the reverse task lies in determination of $C_{\min}$ with the set restriction $E \le E_{\max}$; 3) mini-max method.

The cost-functionality contradiction excludes the best solution by two criteria. Therefore it is useful to join the knowledge and experience of a developer with the automated synthesis of a set of not worse solutions. In this case according to the Pareto's unconditional criterion the set of not worse solutions is determined. Then, the expert selects the final solution from within this set using one conditional criterion of preference and taking into consideration technical perspectives of the device application.

As a result we have received a formal solution of the discrete optimization task for design of digital structures based on microcontrollers. This solution is universal for a wide set of optimization tasks.

ACKNOWLEDGMENT

The work presented in this paper is performed under the project #UKC2-005073-KV-07, "Dynamically reprogrammable network control application processor with Internet capability", supported by the U.S. Civilian Research & Development Foundation.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

ANALYSIS OF THE 2-SUM PROBLEM AND THE SPECTRAL ALGORITHM

Alexander Kolomiychuk

Applied Mathematics Department, Odessa National Polytechnic University,
1, Shevchenko avenue, Odessa, 65044, Ukraine
E-mail: kolomiychuk@gmail.com

Abstract: *This paper presents the analysis of the 2-sum problem and the spectral algorithm. The spectral algorithm was proposed by Barnard, Pothen and Simon in [1]; its heuristic properties have been advocated by George and Pothen in [4] by formulation of the 2-sum problem as a Quadratic Assignment Problem. In contrast to that analysis another approach is proposed: permutations are considered as vectors of Euclidian space. This approach enables one to prove the bound results originally obtained in [4] in an easier way. The geometry of permutations is considered in order to explain what are 'good' and 'pathological' situations for the spectral algorithm. Upper bounds for approximate solutions generated by the spectral algorithm are proved. The results of numerical computations on (graphs of) large sparse matrices from real-world applications are presented to support the obtained results and illustrate considerations related to the 'pathological' cases.*

Keywords: *2-sum problem, Spectral Algorithm, Fiedler vector, Laplacian of a graph, graph layout problems, sparse matrices, spectral graph theory.*

1. INTRODUCTION

The 2-sum problem (to be defined in section 2) is one of the graph enumeration problems (also known as graph layout problems). Besides its purely theoretical interest, it was used as an approximation for large-scale graph labeling problems, that are important in numerical computations using large sparse matrices. Barnard, Pothen and Simon proposed spectral algorithm for envelope reduction of sparse matrices [1]. In fact, their algorithm yields an approximate solution of a 2-sum problem, which itself is used as an approximation for reducing envelope and envelope-related parameters of large sparse matrices. The spectral approach has also been used in the graph partitioning for finding the pseudoperipheral nodes of graphs, and other similar or related problems (e.g. [5,8]).

Fiedler studied the properties of the second Laplacian eigenvalue and eigenvector (also called Fiedler vector, i.e. eigenvector of Laplacian of a connected graph that corresponds to the second smallest eigenvalue). He observed that the differences between the components of this eigenvector are an approximate measure of the distance between the vertices [2,3]. Juvan and Mohar advocated the use of this eigenvector to compute bandwidth and p-sum (more general problem then 2-sum) reducing orderings [6]. [7] is a survey of the applications of Laplacian spectra to different combinatorial problems.

An approximate solution computed by spectral algorithm, generally, is only a heuristic approximation as the 2-sum problem was proved to be NP-complete [4]. Nevertheless, solutions generated by spectral algorithm are known to be quite good for minimizing the envelope of many real-world large sparse matrices [1].

In the paper [4], which is companion-paper for [1], George and Pothen provide analysis of spectral algorithm via Quadratic Assignment Problem (QAP). They formulate 2-sum problem as QAP, and analyze it utilizing permutation matrices. In this paper a different approach is presented (section 3) which instead considers permutations as Euclidian space vectors. Additionally, we present analysis of how effective or ineffective spectral algorithm can be for minimizing the 2-sum problem itself (not necessarily for envelope-reduction); as well as considerations leading to 'pathological' cases are provided. In section 4, computational results are presented to illustrate and test ideas from section 3 on Laplacians of graphs associated with real-world large sparse matrices.

The following notation is used throughout the paper: the underscore symbol _ indicates that the underlined letter is a column-vector (for example

$\underline{x}$); $\underline{u}$ is a column-vector all of whose components are 1.

If otherwise not mentioned, all vectors considered are non-zero column-vectors with dimension n; $n \geq 3$; some of the following considerations may not be correct for degenerate dimensions $n = 1$ and $n = 2$. The distance between vectors is as per Euclidian norm.

2. 2-SUM PROBLEM AND SPECTRAL ALGORITHM

In this section spectral algorithm is presented mainly following Barnard, Pothen and Simon's original work [1]; and as advocated by George and Pothen in [4].

We will consider connected and undirected graphs. Any graph of this type is associated with a symmetric matrix according to the following rule: there is an edge in the associated graph between vertices i and j if and only if the element a_{ij} of matrix A is nonzero. We will consider only matrices that have connected associated graphs.

Let us denote the column indices of the nonzero elements in the lower triangular part of the i-th row:

$$row(i) = \left\{ j : a_{ij} \neq 0 \ \ and \ \ 1 \leq j \leq i \right\}.$$

In these terms the 2-sum problem can be formulated as follows:

$$\sigma_2^2 = \sum_{i=1}^{n} \sum_{j \in row(i)} (i-j)^2 \qquad (2.1)$$

i.e. the sum of squares of the differences between the vertices numbers which are in pairs that share a common edge in the corresponding associated graph. Hereinafter, the 2-sum problem will denote the problem of finding enumeration of graph vertices that minimizes (2.1). Obviously, the set of possible solutions to this problem is a set of permutations; let P denote the set of all permutations with length n.

The Laplacian matrix Q of an undirected graph G is the $n \times n$ matrix $(D-B)$ where D is the diagonal degree matrix, and B is the adjacency matrix of G. Laplacian matrix Q could be defined directly in terms of matrix A as:

$$q_{ij} = \begin{cases} -1, & i \neq j \ \ and \ a_{ij} \neq 0 \\ 0, & i \neq j \ \ and \ \ a_{ij} = 0 \\ -\sum_{\substack{j=1 \\ j \neq i}}^{n} q_{ij}, & i = j \end{cases} \qquad (2.2)$$

The eigenvalues of Q are the Laplacian eigenvalues of G, and we list them as $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. An eigenvector corresponding to λ_k will be denoted by $\underline{x}_k$, and called a k-th eigenvector of Q. It is well-known that Q is a singular M-matrix, and hence its eigenvalues are nonnegative. Thus, $\lambda_1 = 0$, and the corresponding eigenvector is any nonzero constant vector. If G is connected, then Q is irreducible, and $\lambda_2 > 0$ (in the following we suppose that G is connected, otherwise 2-sum problem could be solved separately on each connected component).

Note the following properties of matrix Q that are used later, but not separately specified there:

1. $\underline{x}^T Q \underline{y} = \underline{y}^T Q \underline{x}$ (as Q is symmetrical);

2. $Q\underline{u} = \underline{u}^T Q = 0$ (by the definition of Q).

The idea is to consider the related 2-sum problem, and then show that a second Laplacian eigenvector $\underline{x}_2$ solves a continuous relaxation of the problem. Then it will be proved that the permutation vector, computed by spectral algorithm, is the closest vector among the permutation vectors to eigenvector $\underline{x}_2$. In [1] the following considerations have been proposed.

For odd n, let T denote the set of vectors whose components are permutations of $\{-(n-1)/2, \ldots, -1, 0, 1, \ldots, (n-1)/2\}$. For even n, let T denote vectors that are permutations of $\{-n/2, \ldots, -1, +1, \ldots, n/2\}$. Consider the 2-sum of a symmetric matrix A, defined with respect to vectors in T:

$$\sigma_2^2 = \frac{1}{2} \sum_{a_{ij} \neq 0} (t_i - t_j)^2 \to \min, \ \ t \in T \qquad (2.3)$$

Note that any $\underline{t} \in T$ satisfies $\underline{t}^T u = 0$ and $l \equiv \underline{t}^T \underline{t} = (n/12)(n^2-1)$ for odd n, and $l \equiv \underline{t}^T \underline{t} = (n/12)(n+1)(n+2)$ for even n. Given a vector $\underline{x} \in R^n$, let us define permutation vector t *induced* by x by the rule $t_i \leq t_j$ if and only if $x_i \leq x_j$. Hence, to obtain a continuous relaxation of the discrete problem, consider the set of vectors $x \in R^n$ satisfying $x \neq 0, x^T u = 0$ and $x^T x = l$. This becomes the continuous optimization problem:

$$\frac{1}{2} \min_{x \in X} \sum_{a_{ij} \neq 0} (x_i - x_j)^2 =$$

$$= \min_{x \in X} \left(\sum_{i=1}^{n} d_i x_i^2 - 2 \sum_{\substack{j<i \\ a_{ij} \neq 0}} x_i x_j \right)$$

$$= \min_{x \in X} x^T Dx - x^T Bx = \min_{x \in X} x^T Qx$$
$$= \lambda_2 x_2^T x_2 = \lambda_2 l. \quad (2.4)$$

Hence, a second Laplacian eigenvector $\underline{x}_2$ solves the continuous approximation of the 2-sum problem.

Spectral Algorithm (for 2-sum problem):

Step 1. Given a graph G, generate the Laplacian matrix Q.

Step 2. Compute a second eigenvector $\underline{x}_2$ of Q.

Step 3. Generate a permutation induced by $\underline{x}_2$.

Consider inducing more generally. Vectors $\underline{x}$ and $\underline{y}$ *are in an inducing relation*, (or equivalently: $\underline{x}$ induces $\underline{y}$, or $\underline{y}$ induces $\underline{x}$) if $y_i \le y_j$ if and only if $x_i \le x_j$; and at the same time, there are no additional restrictions on vectors introduced by this inducing relation. Obviously, inducing relation introduces reflexive, symmetric and transitive properties i.e. it is the equivalence relation on a set of vectors, none of which have identical components.

Lemma 2.1. Let there be vector $\underline{y}$ and a set Z that consists of $n!$ vectors, representing all possible permutations of any fixed set of components.

Then, vector $\underline{z} \in Z$, induced by vector $\underline{y}$, is closest to $\underline{y}$ among all vectors of set Z. If vector $\underline{y}$ induces several vectors from Z, then all of them will be at the same distance from $\underline{y}$.

Proof. By the definition of Euclidian norm:

$$\left\| \underline{y} - \underline{z} \right\| = \sqrt{\underline{y}^T \underline{y} + \underline{z}^T \underline{z} - 2\underline{y}^T \underline{z}} \quad (2.5)$$

Since vectors from Z differ only in their component permutation, and vector $\underline{y}$ is fixed, scalar products $\underline{y}^T \underline{y}$ and $\underline{z}^T \underline{z}$ are positive constants, and $\underline{z}^T \underline{z}$ does not depend on the choice of $\underline{z}$. Therefore, in order to minimize the distance $\left\| \underline{y} - \underline{z} \right\|$ vector $\underline{z}$ has to maximize scalar product $\underline{y}^T \underline{z}$. Let us prove that this is true only if vector $\underline{z}$ is induced by vector $\underline{y}$. By contradiction: take a vector $\hat{\underline{z}} \in Z$, which is not in inducing relation with vector $\underline{y}$. Therefore, at least for the two components $\hat{z}_i$ and $\hat{z}_j$ of this vector, it is true that $\hat{z}_i < \hat{z}_j$ i.e. $\hat{z}_j = \hat{z}_i + a$, $a > 0$ and $y_i > y_j$. Now, scalar product $\underline{y}^T \hat{\underline{z}}$ increases if components $\hat{z}_i$ and $\hat{z}_j$ are swapped, while the other components remain in their places:

$$y_i \hat{z}_i + y_j \hat{z}_j = y_i(\hat{z}_j - a) + y_j(\hat{z}_i + a) =$$
$$= y_i \hat{z}_j + y_j \hat{z}_i + a(y_j - y_i) < y_i \hat{z}_j + y_j \hat{z}_i$$

since $a > 0$ and $(y_j - y_i) < 0$.

Thus, vector $\hat{\underline{z}}$ does not maximize the scalar product $\underline{y}^T \underline{z}$; and, hence, proof by contradiction is completed.

One should note that if vector $\underline{y}$ induces several vectors from Z, then all of them (by definition of inducing relation) have the same value of scalar product with $\underline{y}$ i.e. will be at the same distance from it.

The lemma proven is a slightly more general analogue of the corresponding theorem from [1].

The set of permutations P contains vectors that can be obtained from each other by all-possible component permutations. Therefore, according to the proven lemma, the permutation induced by vector $\underline{x}_2$ is the closest to it among all permutations.

The heuristic idea of choosing the permutation closest to the accurate solution naturally has to be such that, given the close position of the permutation chosen to the accurate solution, the increase of the cost function (quadratic form of matrix Q) remains small enough.

3. ANALYSIS

This section presents the analysis of the 2-sum problem. It is performed by the author of this paper in a quite different way than those in papers [2] and [1] for exploring and justifying spectral algorithm.

Let us consider permutations as vectors in n-dimension Euclidian space; we will denote the set of permutations as P. The following identities for any $p \in P$ are elementary:

$$1 + 2 + \ldots + n = \underline{p}^T \underline{u} = \frac{n(n+1)}{2},$$
$$1^2 + 2^2 + \ldots n^2 = \underline{p}^T \underline{p} = \frac{n(n+1)(2n+1)}{6} \quad (3.1)$$

Let V be a set of vectors $\underline{v} \in \mathrm{R}^n$:

$$\begin{cases} \underline{v}^T \underline{u} = \dfrac{n(n+1)}{2} \\ \underline{v}^T \underline{v} = \dfrac{n(n+1)(2n+1)}{6} \end{cases} \quad (3.2)$$

It is obvious that $P \subset V$.

For further analysis of the 2-sum problem in the relaxed form we will use a set X of admissible solutions. X is a set of non-zero vectors $\underline{x} \in \mathrm{R}^n$,

satisfying the following conditions:

$$\begin{cases} \underline{x}^T \underline{x} = 1 \\ \underline{x}^T \underline{u} = 0 \end{cases} \quad (3.3)$$

Note that all our further analyses can be similarly performed by considering the first constraint in a generalized form: $\underline{x}^T \underline{x} = l$ $(l \in \mathrm{R}, l \neq 0)$. We suppose $l = 1$ for the purpose of laconic formulation.

Consider a mapping of set V onto set X:

$$\underline{x} = \alpha(\underline{v} + \beta \underline{u}), \quad \alpha, \beta \in \mathrm{R}, \alpha \neq 0 \quad (3.4)$$

By considering conditions (3.2) and (3.3) we can define constants α and β:

$$\begin{aligned} \underline{x}^T \underline{u} &= \alpha(\underline{v} + \beta \underline{u})^T \underline{u} = \\ &= \alpha\left(\frac{n(n+1)}{2} + \beta n\right) = 0 \end{aligned}, \quad (3.5)$$

and since $\alpha \neq 0$, we get $\beta = -\frac{n+1}{2}$.

Similarly we obtain:

$$\begin{aligned} &\underline{x}^T \underline{x} = \alpha^2 (\underline{v} + \beta \underline{u})^T (\underline{v} + \beta \underline{u}) = 1 \\ &\Rightarrow \alpha = \sqrt{\frac{12}{n(n^2-1)}} \end{aligned} \quad (3.6)$$

Note that the mapping (3.4) is an affine transformation that compresses and shifts vectors in V.

Now, when expressing $\underline{v}$ from (3.4), we can show the mapping inverse to (3.4) is in the following form:

$$\underline{v} = \frac{1}{\alpha}(\underline{x} - \alpha\beta \underline{u}) = \alpha'(\underline{x} + \beta' \underline{u}), \quad (3.7)$$

$$\begin{aligned} \alpha' &= \frac{1}{\alpha} = \sqrt{\frac{n(n^2-1)}{12}}, \\ \beta' &= -\alpha\beta = \sqrt{\frac{3(n+1)}{n(n-1)}} \end{aligned} \quad (3.8)$$

Let us denote as $\tilde{X}$ a set of images of the transformation (3.4), the preimages of which are permutations from P. Since $P \subset V$, then $\tilde{X} \subset X$. Let the elements of set $\tilde{X}$ be called *representatives* of the corresponding permutations.

Thus, mappings (3.4) and (3.7) define the one-to-one dependence between sets V and X (and between their subsets P and $\tilde{X}$).

Lemma 3.1. $\breve{\underline{x}}^T Q \breve{\underline{x}} > \hat{\underline{x}}^T Q \hat{\underline{x}}$ if and only if $\breve{\underline{v}}^T Q \breve{\underline{v}} > \hat{\underline{v}}^T Q \hat{\underline{v}}$; and also $\breve{\underline{x}}^T Q \breve{\underline{x}} = \hat{\underline{x}}^T Q \hat{\underline{x}}$ if and only if $\breve{\underline{v}}^T Q \breve{\underline{v}} = \hat{\underline{v}}^T Q \hat{\underline{v}}$,

where $\breve{\underline{x}} = \alpha(\breve{\underline{v}} + \beta \underline{u})$; $\hat{\underline{x}} = \alpha(\hat{\underline{v}} + \beta \underline{u})$; $\breve{\underline{x}}, \hat{\underline{x}} \in X$; $\breve{\underline{v}}, \hat{\underline{v}} \in V$.

Proof follows from (3.4), (3.7) and the identity:

$$(\alpha(\underline{y} + \beta \underline{u}))^T Q(\alpha(\underline{y} + \beta \underline{u})) = \alpha^2 \underline{y}^T Q \underline{y}.$$

In view of Lemma 3.1 we can analyze the discrete 2-sum problem not on set P, but on set $\tilde{X}$ and, correspondingly, examine its continuous variant on set X:

$$\begin{cases} \underline{x}^T Q \underline{x} \to \min \\ \underline{x} \in X \end{cases} \quad (3.9)$$

This problem is similar to (2.3) and, according to the well-known theorem of Courant-Fischer, the minimum is reached on vector $\underline{x}_2$ i.e. the second eigenvector of matrix Q; and maximum on vector $\underline{x}_n$:

$$\min_{\underline{x} \in X} \{\underline{x}^T Q \underline{x}\} = \lambda_2; \quad \max_{\underline{x} \in X} \{\underline{x}^T Q \underline{x}\} = \lambda_n \quad (3.10)$$

This can also be shown in another way: by using the Lagrange method for finding the conditional extremum of a multivariable function.

According to the above choice of permutation induced by vector $\underline{x}_2$, on Step 3 of spectral algorithm, the later can be presented as two transfers:

1) Transfer from vector $\underline{x}_2$ to vector $\tilde{\underline{x}}^*$ from $\tilde{X}$, which is induced by $\underline{x}_2$ (i.e. is the closest to it in $\tilde{X}$ according to Lemma 2.1). Thus, an approximate solution of 2-sum problem on set $\tilde{X}$ is obtained.

2) Transfer from vector $\tilde{\underline{x}}^*$ to the permutation it represents, according to (3.7), i.e. approximate solution of the 2-sum problem on set P is obtained.

The above sets and representations allow us to derive a simple proof of the following theorem, which was initially proved in [4] but using another approach.

Theorem 3.1. The following upper and lower bounds hold for the 2-sum problem:

$$\begin{aligned} &(1/12)\lambda_2 n(n-1)(n+1) \le \sigma_2^2 \le \\ &\le (1/12)\lambda_n n(n-1)(n+1) \end{aligned}.$$

Proof. Let us prove the lower bound. Since $\tilde{X} \subset X$:

$$\min_{\tilde{\underline{x}} \in \tilde{X}} \{\tilde{\underline{x}}^T Q \tilde{\underline{x}}\} \ge \min_{\underline{x} \in X} \{\underline{x}^T Q \underline{x}\},$$

and taking into account (3.10)

$$\min_{\tilde{\underline{x}} \in \tilde{X}} \{\tilde{\underline{x}}^T Q \tilde{\underline{x}}\} \ge \lambda_2.$$

In view of Lemma 3.1 and Corollary 1 it is true

that:

$$\min_{\underline{p}\in P}\{\underline{p}^T Q\,\underline{p}\} = (\alpha')^2 \min_{\tilde{\underline{x}}\in\tilde{X}}\{\tilde{\underline{x}}^T Q\tilde{\underline{x}}\} \geq (\alpha')^2\lambda_2 .$$

The substitution of value α' from (3.8) concludes the proof of the lower bound for σ_2^2. The proof of the upper-bound is analogous.

For clearer illustration of the following argumentation let us present a geometric interpretation of sets X and $\tilde{X}$. The equation $\underline{x}^T\underline{x}=1$ from (3.3) defines an n-dimensional hypersphere with radius 1 at the point of origin; and the equation $\underline{x}^T\underline{u}=0$ defines an n-dimensional hyperplane that intersects the point of origin and is orthogonal to vector $c\underline{u}, c\in \mathrm{R}$. Thus, set X is an $(n-1)$-dimensional hypersphere that is obtained from the intersection of the n-dimensional hypersphere with the n-dimensional hyperplane that intersects its center. Fig.1 is an illustration of the case $n=3$. Hereinafter hypersphere will simply be called a sphere.

A set $\tilde{X}$ represents an $(n-1)$-dimensional polyhedron, the vertices of which are located on sphere X. It is easy to show that the permutations closest to each other, are those obtained from each other by simply permuting a pair of their components that differ by value 1; for example, $(1,2,3,4)$ and $(1,3,2,4)$. Similarly, the representatives of the permutations closest to each other are obtained from each other by permuting one pair of components that differ by value α.

When $n=3$ the permutations' representatives are the vertices of the planar hexagon shown in Fig.2 (this hexagon is, in fact, inscribed in the shaded circle shown in Fig.1); near the hexagon's vertices the corresponding permutations are indicated.

When $n=4$ all permutations are located on the $n-1=3$ dimensional sphere, illustrated in Fig.3. The pairs of permutations closest to each other are connected, with the edges forming the polyhedron. Note that Fig.3 is not a schematic picture, but a 3-dimensional visualization made using Matlab.

To calculate the 3-dimensional coordinates, all 4!=24 permutations were first mapped into set $\tilde{X}$ according to (3.4). Then in plane $\underline{x}^T\underline{u}=0$, the orthonormalized basis was chosen comprising 3 vectors; and in this basis, the permutations are represented in 3 dimensions (the fourth component of all permutations is 0).

Note that since the mapping (3.7) is an affine transform, the above description of the permutations' representatives characterizes the geometry of the permutations themselves.

Now, let us analyze possible increase of cost function value in problem (3.9) when transferring from vector $\underline{x}_2$ to its nearest vector $\tilde{\underline{x}}^*\in\tilde{X}$. It should be noted that the distance between vectors on sphere X is, in fact, determined by the scalar product of these vectors because

$$\left\|\underline{x}-\underline{y}\right\| = \sqrt{(\underline{x}-\underline{y})^T(\underline{x}-\underline{y})} = \sqrt{2(1-\underline{x}^T\underline{y})},$$

where $\underline{x},\underline{y}\in X$. According to (3.3) $\underline{x}\in X \Leftrightarrow -\underline{x}\in X$, and considering the nearest vectors, it is assumed that the scalar product of any pair of vectors belongs to $[0,1]$ (when transferring from $\underline{x}$ to $-\underline{x}$, the value of the quadratic form of matrix Q does not change).

If $\tilde{\underline{x}}^*\neq\underline{x}_2$, then vector $\tilde{\underline{x}}^*$ can be considered as located on the arc of sphere X, which connects $\underline{x}_2$ and some orthogonal to it vector $\underline{z}\in X$, $\underline{z}^T\underline{x}_2=0$ (see Fig. 4) :

$$\tilde{\underline{x}}^* = \frac{a\underline{x}_2+(1-a)\underline{z}}{\left\|a\underline{x}_2+(1-a)\underline{z}\right\|},\quad a\in[0,1] \qquad (3.11)$$

Let us call vector $\underline{z}$ a *spherical orthogonal continuation* of vector $\underline{x}_2$ through vector $\tilde{\underline{x}}^*$.

From (3.11) by the definition

$$(\underline{x}_2^T\tilde{\underline{x}}^*)\left\|a\underline{x}_2+(1-a)\underline{z}\right\| = a$$

and so

$$(\underline{x}_2^T\tilde{\underline{x}}^*)^2((a\underline{x}_2+(1-a)\underline{z})^T(a\underline{x}_2+(1-a)\underline{z})) = a^2$$

.

Therefore we can obtain the following equations:

$$\begin{aligned}(\underline{x}_2^T\tilde{\underline{x}}^*)^2 &= \frac{a^2}{(a^2+(1-a)^2)};\\ 1-(\underline{x}_2^T\tilde{\underline{x}}^*)^2 &= \frac{(1-a)^2}{(a^2+(1-a)^2)}\end{aligned} \qquad (3.12)$$

Now, the cost function value of the problem (3.9) on vector $\tilde{\underline{x}}^*$ can be represented as follows using (3.12):

$$\begin{aligned}&\tilde{\underline{x}}^{*T}Q\tilde{\underline{x}}^* =\\ &= \frac{a^2\lambda_2+(1-a)^2\underline{z}^TQ\underline{z}}{a^2+(1-a)^2} =\\ &= (\underline{x}_2^T\tilde{\underline{x}}^*)^2\lambda_2+(1-(\underline{x}_2^T\tilde{\underline{x}}^*)^2)\underline{z}^TQ\underline{z}\end{aligned} \qquad (3.13)$$

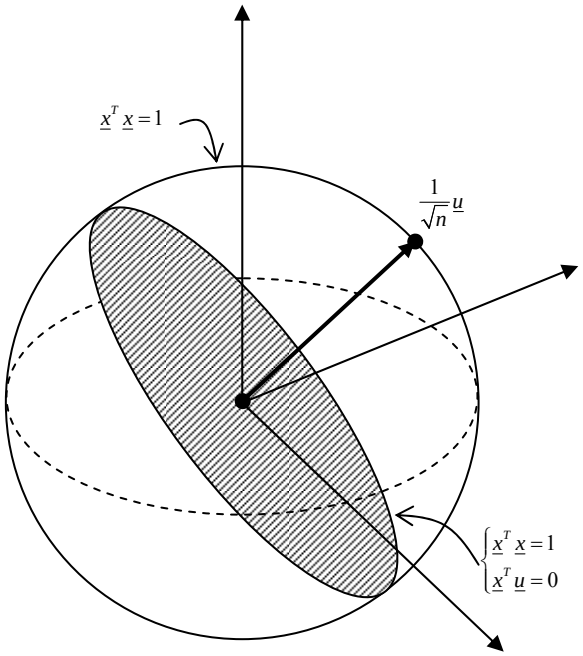


Fig. 1 – Geometry of the set X

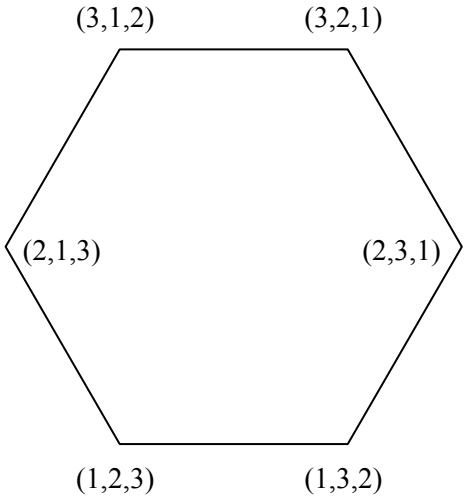


Fig. 2 – Geometry of permutation, n=3

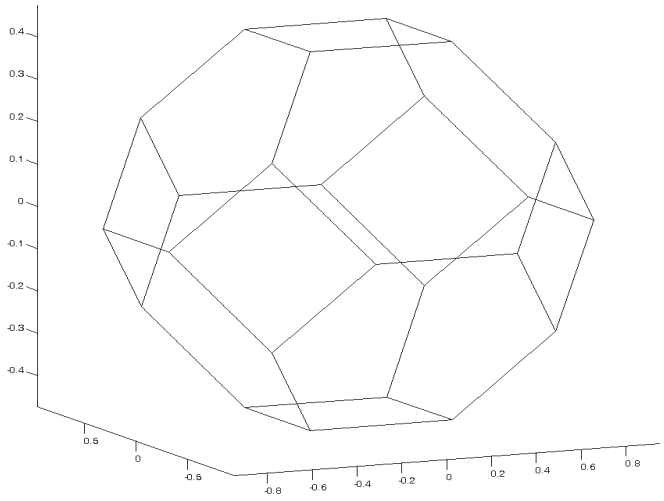


Fig. 3. Geometry of permutations, n=4

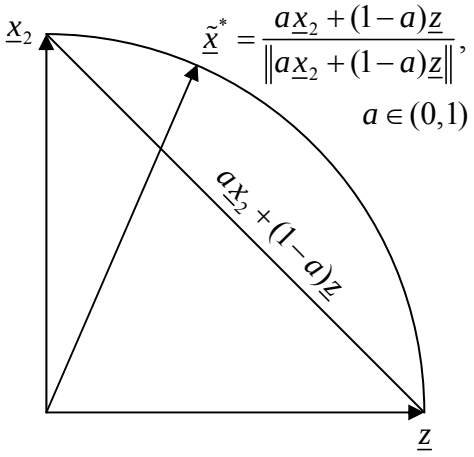


Fig. 4. Spherical combination

As easily seen from (3.13), it follows that the quadratic form of matrix Q strictly increases along the arc connecting $\underline{x}_2$ and $\underline{z}$; and with the fixed distance from $\underline{x}_2$ to $\tilde{\underline{x}}^*$, its value is defined by the value $\underline{z}^T Q\underline{z}$.

Thus, the increase of cost function value in problem (3.9), when transferring from $\underline{x}_2$ to $\tilde{\underline{x}}^*$, depends both on the distance from $\underline{x}_2$ to $\tilde{\underline{x}}^*$ (the bigger the distance, the bigger the function value), and on the direction in which such a transfer is performed (i.e. on cost function value on the spherical orthogonal continuation of vector $\underline{x}_2$ through vector $\tilde{\underline{x}}^*$).

Matrix Q is a normal matrix (because it is symmetric), so there is an orthogonal system formed by its eigenvectors. Setting $\underline{x}_1 = (1/\sqrt{n})\underline{u}$ an orthonormalized system of eigenvectors $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$, $\underline{x}_i^T\underline{x}_j = 0 \quad (i \neq j)$ is obtained, that

is a basis of the space R^n. In this basis, vector $\tilde{\underline{x}}^*$ is presented as follows:

$$\tilde{\underline{x}}^* = a_2 \underline{x}_2 + a_3 \underline{x}_3 + \ldots + a_n \underline{x}_n \qquad (3.14)$$

because $\tilde{\underline{x}}^{*T} \underline{u} = 0$, and

$$\begin{aligned} &\tilde{\underline{x}}^{*T} \tilde{\underline{x}}^* = a_2^2 + a_3^2 + \ldots + a_n^2 = 1; \\ &(\underline{x}_i^T \tilde{\underline{x}}^*)^2 = a_i^2 \quad (2 \le i \le n) \end{aligned} \qquad (3.15)$$

From (3.14) and (3.15) it follows that

$$\begin{aligned} &\tilde{\underline{x}}^{*T} Q \tilde{\underline{x}}^* = \\ &= a_2^2 \lambda_2 + a_3^2 \lambda_3 + \ldots + a_n^2 \lambda_n \le \\ &\le a_2^2 \lambda_2 + (a_3^2 + \ldots + a_n^2) \lambda_n = \\ &= a_2^2 \lambda_2 + (1 - a_2^2) \lambda_n \end{aligned} \qquad (3.16)$$

Note that we now come to the same result as in (3.13) if we suppose $\underline{z} = \arg\max_{\underline{x} \in X} \{\underline{x}^T Q \underline{x}\} = \underline{x}_n$.

Lemma 3.2. The following bound is true:

$$\begin{aligned} &(\underline{x}_2^T \tilde{\underline{x}}^*)^2 \lambda_2 + (1 - (\underline{x}_2^T \tilde{\underline{x}}^*)^2) \lambda_3 \le \tilde{\underline{x}}^{*T} Q \tilde{\underline{x}}^* \le \\ &\le (\underline{x}_2^T \tilde{\underline{x}}^*)^2 \lambda_2 + (1 - (\underline{x}_2^T \tilde{\underline{x}}^*)^2) \lambda_n \end{aligned}$$

where $\tilde{\underline{x}}^*$ is the nearest vector to $\underline{x}_2$ from $\tilde{X}$ (i.e. $\underline{x}_2$ induces $\tilde{\underline{x}}^*$).

Proof of the upper bound follows from (3.16) or from (3.13) using (3.10).

The lower bound follows from (3.13) if we suppose $\underline{z} = \arg\min_{\underline{x} \in X} \{\underline{x}^T Q \underline{x}\}$ with the additional restriction $\underline{x}^T \underline{x}_2 = 0$; in this case $\underline{z} = \underline{x}_3$ (as in (3.10) it follows from the Courant-Fischer theorem), and then $\underline{z}^T Q \underline{z} = \lambda_3$. □

Thus, the 'worst' transfer direction from $\underline{x}_2$ to $\tilde{\underline{x}}^*$ is an arc connecting $\underline{x}_2$ and $\underline{x}_n$, although it can be compensated by the nearness of $\tilde{\underline{x}}^*$ to $\underline{x}_2$.

Now, from the cost function decomposition in (3.16) it is clear that even after computing all eigenvectors/eigenvalues of the Laplacian Q, the problem of choosing a vector $\tilde{\underline{x}} \in \tilde{X}$, (even with guaranteed approximation) that minimizes $\tilde{\underline{x}}^T Q \tilde{\underline{x}}$ remains hard because the cost function depends on the squares of coefficients of vector $\tilde{\underline{x}}$ decomposition in the basis of Q eigenvectors.

Besides that, the computation of all eigenvectors/eigenvalues of large matrices is almost impossible in practice. Given this point of view, the heuristic idea of spectral algorithm consists in choosing $\tilde{\underline{x}}^*$, which maximizes $(\underline{x}_2^T \tilde{\underline{x}}^*)^2 = a_2^2$; owing to which one can expect a decrease in the coefficients corresponding to the bigger eigenvalues, and the value of the cost function will most likely be not much bigger than $\underline{x}_2^T Q \underline{x}_2 = \lambda_2$.

It is interesting to determine how far vectors $\underline{x} \in X$ and $\tilde{\underline{x}} \in \tilde{X}$, which are in inducing relation, can be distanced from each other. The following theorem answers this question.

Theorem 3.2. Let $\tilde{\underline{x}} \in \tilde{X}$ be induced by vector $\underline{x} \in X$ which has n_1 negative and n_2 positive components ($n_1 + n_2 = n$, and zero components are considered together with either negative or positive components). Then the following bound is true:

$$\underline{x}^T \tilde{\underline{x}} \ge \sqrt{\frac{3 n_1 n_2}{(n^2 - 1)}},$$

and equality is reached on the unique vector that has all positive components equal to each other and which stand on the same places as the positive components of vector $\underline{x}$; and where all negative components are equal to each other and stand on the same places as the negative components of vector $\underline{x}$.

The theorem has been proved by the author of this paper but the proof is too long and technical to present it here in full. The simplified scheme of the proof is the following:

1. Consider the theorem as the statement about the unique solution of the constrained minimization problem with the cost function $\underline{x}^T \tilde{\underline{x}}$.
2. Prove that the minimum exists.
3. Show that the minimum is unique.
4. Show that only the vector described in the second part of the theorem could be the minimum point.

From the geometrical point of view, the vectors which consist only of equal to each other positive and equal to each other negative components, pass through the 'centers' of the edges of the polyhedron of permutations representatives. So when $n = 4$ (see Fig. 3) the vectors $(-\sqrt{3/4}, \sqrt{1/12}, \sqrt{1/12}, \sqrt{1/12})$, $(-\sqrt{1/12}, -\sqrt{1/12}, -\sqrt{1/12}, \sqrt{3/4})$ and $(-\sqrt{1/4}, -\sqrt{1/4}, \sqrt{1/4}, \sqrt{1/4})$ pass through the centers of two hexagons and one quadrangle that share one vertex which is representative of unit permutation $\underline{p}_1 = (1, 2, 3, 4)$.

Evidently, the value of the bound we obtained in Theorem 3.2 reaches its maximum value when $n_1 = n_2 = n/2$ (approximately for odd n), and its minimum on the points $\{n_1 = 1, n_2 = n - 1\}$ and

$\{n_1 = n-1, n_2 = 1\}$. That is why with unknown n_1 and n_2 the following is true

Corollary 1 from Theorem 3.2. Let $\tilde{\underline{x}} \in \tilde{X}$ be induced by vector $\underline{x} \in X$, then

$$\underline{x}^T \tilde{\underline{x}} \geq \sqrt{\frac{3}{n+1}}.$$

Asymptotical behavior of the bound when $n \to \infty$ essentially depends on the ratio of n_1 and n_2:

If $n_1 = n_2 = \frac{n}{2}$,

then $\lim_{n\to\infty} \sqrt{\frac{3n_1 n_2}{(n^2-1)}} = \lim_{n\to\infty} \sqrt{\frac{3n^2}{4(n^2-1)}} = \sqrt{\frac{3}{4}}$,

if $n_1 = 1, n_2 = n-1$ or $n_1 = n-1, n_2 = 1$ then

$$\lim_{n\to\infty} \sqrt{\frac{3n_1 n_2}{(n^2-1)}} = \lim_{n\to\infty} \sqrt{\frac{(n-1)}{(n^2-1)}} = 0$$

i.e. in the 'worst case' (for example, if $n_1 = 1, n_2 = n-1$) vectors $\underline{x}^T$ and $\tilde{\underline{x}}$ are 'asymptotically orthogonal'.

Applying the Theorem 3.2 to $\underline{x}_2$ and $\tilde{\underline{x}}^*$ we obtain: $\underline{x}_2^T \tilde{\underline{x}}^* \geq \sqrt{\frac{3n_1 n_2}{(n^2-1)}}$, where n_1 and n_2 correspondingly denote the number of positive and negative components in vector $\underline{x}_2$. Substituting this bound in the inequality of Lemma 3.2 we obtain:

$$\tilde{\underline{x}}^{*T} Q \tilde{\underline{x}}^* \leq \frac{3n_1 n_2}{(n^2-1)}\lambda_2 + \left(1 - \frac{3n_1 n_2}{(n^2-1)}\right)\lambda_n =$$

$$= \frac{1}{n^2-1}\left(3n_1 n_2 \lambda_2 + (n^2 - 3n_1 n_2 - 1)\lambda_n\right) \quad (3.21)$$

Equality can be reached if in vector $\underline{x}_2$ all the negative components equal c and all the positive components equal k. Then all permutations where numbers $1, 2, \ldots, n_1$ stand on the same places as in c in $\underline{x}_2$, and numbers $n_1 + 1, \ldots, n$ stand on the same places as k in $\underline{x}_2$ are closest to $\underline{x}_2$ in P (in all there are $n_1! n_2!$ such permutations), and their representatives are closest to $\underline{x}_2$ in $\tilde{X}$. In this case spectral algorithm can choose any of these permutations and in the 'worst case' the chosen permutation will be located on the arc connecting $\underline{x}_2$ and $\underline{x}_n$.

The asymptotical behavior of the bound (3.21) also depends substantially on the relation between n_1 and n_2:

If $n_1 = n_2 = \frac{n}{2}$,

then $\lim_{n\to\infty} \tilde{\underline{x}}^{*T} Q \tilde{\underline{x}}^* = \frac{3}{4}\lambda_2 + \frac{1}{4}\lambda_n$,

if $n_1 = 1, n_2 = n-1$ or $n_1 = n-1, n_2 = 1$,

then $\lim_{n\to\infty} \tilde{\underline{x}}^{*T} Q \tilde{\underline{x}}^* = \lambda_n$.

Nevertheless, the above bounds are usually not reached on the graphs of big sparse matrices from real-world applications (as described in Section 4) because the above described 'pathological' situations are not realized. To the contrary, the value of the quadratic form $\tilde{\underline{x}}^{*T} Q \tilde{\underline{x}}^*$ usually is not much bigger than λ_2 while λ_n is several orders bigger than λ_2.

Nevertheless, strictly speaking, with a fixed n we have only a finite number of usual Laplacians, but there are infinitely many weighted ones, whose components can differ greatly. The above analysis can be applied to any Laplacian, including a weighted Laplacian. The obtained bounds still hold without additional restrictions being imposed on a graph, e.g. the degrees of its vertices need not be bounded.

4. COMPUTATIONAL RESULTS

This section lists the results obtained for examining and illustrating the performance of spectral algorithm on Laplacians of real-world large sparse matrices. The matrices used were taken from the Harwell-Boeing and NASA collections that are in Tim Davis' University of Florida Sparse Matrix Collection [9]. These matrices are often used for testing and comparing reordering algorithms for sparse matrices. All computations have been performed in Matlab 7 (Release 14, Service Pack 2); vectors $\underline{x}_2$ and $\underline{x}_n$ were computed using `eigs` function (as defined in Section 2, $\underline{x}_2$ and $\underline{x}_n$ are eigenvectors of the graph's Laplacian corresponding to the 2nd and the largest eigenvalues).

Table 1 presents the following data:

n is a matrix (graph) dimension;

$|E|$ is the number of edges of a graph ($2|E| + n$ is the number of the non-zero matrix elements);

Deg. min / max are minimum and maximum degrees of graph vertices;

pos./neg. values in $\underline{x}_2$ are numbers of positive and negative components in $\underline{x}_2$;

$F(\underline{x}_2)$, $F(\tilde{\underline{x}}^*)$, $F(z)$, $F(\underline{x}_n)$ are values of the Laplacian quadratic forms on the corresponding vectors; vector $\underline{z}$ is the spherical orthogonal

continuation of vector $\underline{x}_2$ through vector $\tilde{\underline{x}}^*$ (these vectors were defined and discussed in Section 3);

$\underline{x}_2^T \tilde{\underline{x}}^*$ is the scalar product of vectors $\underline{x}_2$ and $\tilde{\underline{x}}^*$ that shows the nearness of $\tilde{\underline{x}}^*$ to $\underline{x}_2$.

Table 1. Experimentation with the real-world large sparse matrices (graphs)

Matrix	*n*	$\|E\|$	*Deg. min / max*	*pos./neg. values in* $\underline{x}_2$	$F(\underline{x}_2)$	$F(\tilde{\underline{x}}^*)$	$F(\underline{z})$	$F(\underline{x}_n)$	$\underline{x}_2^T \tilde{\underline{x}}^*$
CAN1054	1,054	5,571	5 / 34	558 / 496	5.93e-2	6.65e-2	8.56e-1	3.57e+1	0.995
CAN1072	1,072	5,686	5 / 34	480 / 592	7.96e-2	8.78e-2	6.11e-1	3.57e+1	0.992
BCSSTK15	1,505	5,406	3 / 34	90 / 1,415	4.24e-2	1.13e+0	1.35e+0	3.54e+1	0.418
NASA1824	1,824	18,692	5 / 41	913 / 911	2.71e-1	3.58e-1	5.21e+0	4.42e+1	0.991
NASA2146	2,146	35,052	13 / 35	1,066 / 1,080	1.35e-1	1.61e-1	1.88e+0	4.77e+1	0.992
NASA2910	2,910	85,693	15/ 174	1,634 / 1,276	1.10e+0	1.74e+0	1.62e+1	1.76e+2	0.978
NASA4704	4,704	50,026	5 / 41	2,438 / 2,266	8.26e-2	9.33e-2	2.36e+0	4.43e+1	0.998
BARTH4	6,019	17,473	3 / 12	3,453 / 2,566	1.77e-3	2.91e-3	2.13e-2	1.34e+1	0.970
BARTH	6,691	19,748	3 / 12	3,354 / 3,337	2.60e-3	2.68e-3	1.47e-1	1.34e+1	1.000
SHUTTLE_EDDY	10,429	46,585	3 / 26	5,389 / 5,040	6.23e-4	2.28e-3	4.86e-2	2.74e+1	0.983
BARTH5	15,606	45,878	3 / 10	6,816 / 8,790	7.70e-4	9.67e-4	5.55e-3	1.17e+1	0.979
BCSSTK30	28,924	1,007,284	3 / 218	15,428 / 13,496	1.95e-2	2.84e-2	5.68e-1	2.22e+2	0.992
BCSSTK32	44,609	985,046	1 / 215	7,079 / 37,530	6.00e-3	2.86e-2	5.15e-2	2.17e+2	0.71

As we can see from Table 1, in most cases the closest representative to $\underline{x}_2$ is very closely located (product $\underline{x}_2^T \tilde{\underline{x}}^*$ is close to 1); and the value of $F(\tilde{\underline{x}}^*)$ is not much bigger than $F(\underline{x}_2)$, at least $F(\tilde{\underline{x}}^*)$ is of the same order as $F(\underline{x}_2)$. At the end of an arc (on vector $\underline{z}$) the value $F(\underline{z})$ is at least one order bigger than the value $F(\tilde{\underline{x}}^*)$ i.e. for $F(\tilde{\underline{x}}^*)$ to be close to $F(\underline{x}_2)$ the nearness of $\tilde{\underline{x}}^*$ to $\underline{x}_2$ is important. At the same time, the value $F(\underline{x}_n)$ is in most cases at least one order bigger than $F(\underline{z})$, which indicates the relatively 'good' direction of shifting along the arc (compare in the 'worst' case $F(\underline{z}) = F(\underline{x}_n)$ i.e. the arc connects $\underline{x}_2$ and $\underline{x}_n$). The example of a very 'good' direction is BARTH5 ($F(\underline{z})$=5.55e-3; $F(\underline{x}_n)$=1.17e+1). It is interesting to note that matrices, where the above situation pertains, have a different size, structure and degree of sparseness. Some of them also have a very dispersed degrees of vertices in the associated graph (for example, in NASA2910 the minimum degree is 15, and maximum 174; with the relatively small dimension $n = 2,910$). Additionally, the close (in some cases almost equal) number of positive and negative components in vector $\underline{x}_2$ is common for these matrices.

A different situation pertains in BCSSTK32, and especially in BCSSTK15. In these $\tilde{\underline{x}}^*$ is located fairly far from $\underline{x}_2$. Because of this the value $F(\tilde{\underline{x}}^*)$ is much bigger than $F(\underline{x}_2)$ and has the same order as $F(\underline{z})$. These two matrices are characterized by a large number of components of one sign and a small number of another sign in $\underline{x}_2$; for BCSSTK15 the ratio is 90 : 1,415.

From the geometrical point of view when the numbers of positive and negative components in $\underline{x}_2$ are close, it is located on the sphere X 'above' 'small' edge of the polyhedron of permutations (see Fig. 3); and that's why we can expect the closest representative of permutation $\tilde{\underline{x}}^*$ to be located very closely ($\underline{x}_2^T \tilde{\underline{x}}^*$ is close to 1). Otherwise, $\underline{x}_2$ is located 'above' 'big' edge and, correspondent vector $\tilde{\underline{x}}^*$ can be located fairly far away ($\underline{x}_2^T \tilde{\underline{x}}^*$ are substantially smaller than 1). In the worst case, $\underline{x}_2$ could be located 'above' the center of the 'biggest' possible edge, and the bound from Corollary 1 of

Theorem 3.2 could be reached.

5. CONCLUSIONS

Spectral algorithm provides a good balance between the computational cost and the accuracy of the approximate solution obtained for NP-complete 2-sum problem. The analysis conducted shows that there are 'pathological' cases for spectral algorithm, and that these are due to the geometrical properties of permutations considered as vectors in Euclidian space. These conclusions are illustrated by the computational results obtained using Laplacians of graphs of big sparse matrices. It should be noted that spectral algorithm can even be used as an effective approximation algorithm for small dimensions graphs; for small graphs it is possible to compute all eigenvalues and eigenvectors, but this additional information would not fundamentally simplify the solution of a 2-sum problem in all cases.

The proposed 'geometrical' approach enables the simple proof of the upper and lower bounds of the 2-sum problem. Potentially it could turn out to be helpful for analysis or design of algorithms for other similar problems like p-sum, graph partitioning, etc.

To the author's knowledge, it is the first time strict (non-trivial) upper bounds are derived for the cost function of a 2-sum problem on an approximate solution provided by spectral algorithm. As the computational experiment shows, these bounds are usually not reached on the graphs of real-world large sparse matrices because the 'pathological' cases are not realized or are realized only partially.

6. AKNOLEDGEMENTS

The author would like to thank Alla A. Kobozeva, associate professor of Applied Mathematics Department, Odessa National Polytechnic University and also the members of her scientific seminar for the support, advices and many helpful discussions of the results presented in this paper.

7. REFERENCES

[1] S. T. Barnard, A. Pothen, H. D. Simon. *A spectral algorithm for envelope reduction of sparse matrices*. Numer. Linear Algebra Appl. 2, No.4, 317-334 (1995).

[2] M. Fiedler. *Algebraic connectivity of graphs*. Czech. Math. J. 23(98), 298-305 (1973).

[3] M. Fiedler. *A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory*. Czech. Math. J. 25(100), 619-633 (1975).

[4] J. A. George, A. Pothen. *An analysis of spectral envelope reduction via quadratic assignment problems*. SIAM J. Matrix Anal. Appl. 18, No.3, 706-732 (1997).

[5] B. Hendrickson, R. Leleand. *An improved spectral graph partitioning algorithm for mapping parallel computations.* SIAM J. Sci. Comput. 16, 452-469 (1995).

[6] M. Juvan, B. Mohar. *Optimal linear labeling and eigenvalues of graphs*. Discr. Appl. Math. 36, 153-168 (1992).

[7] B. Mohar, S. Poljak. *Eigenvalues in combinatorial optimization. Combinatorial and Graph-Theoretical Problems in Linear Algebra.* IMA Volumes in Mathematics and its applications, Vol. 50, Springer-Verlag (1993).

[8] A. Pothen, H. D. Simon, K. P. Liou. *Partitioning sparse matrices with eigenvectors of graphs.* SIAM J. Matrix Anal. Appl. 11, 430-452 (1990).

[9] Tim Davis: University of Florida Sparse Matrix Collection. (http://www.cise.ufl.edu/research/sparse/matrices/)

Alexander Kolomiychuk, *Master of science in applied mathematics. Graduated with honors from Odessa National Polytechnic University (Odessa, Ukraine) in 2005.*

At the present moment: PhD student at the Applied Mathematics Department, Odessa National Polytechnic University.

Research interests: theory and algorithms for combinatorial optimization on graphs.

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

MAXIMAL USING OF SPECIFICS OF SOME BOUNDARY PROBLEMS IN POTENTIAL THEORY AFTER THEIR NUMERICAL ANALYSIS

L.I. Mochurad [1], Y.S. Harasym [2], B.A. Ostudin [3]

[1] Post-Graduate Student of the Department of Computational Mathematics, Ivan Franko National University in Lviv, Mochurad_lesya@ukr.net, http://blues.lnu.edu.ua/ami/kom/mochurad
[2] Senior Lecturer of the Department of Computational Mathematics, Ivan Franko National University in Lviv, garasym@yahoo.com, http://blues.lnu.edu.ua/ami/kom/garasym
[3] Associate Professor of the Department of Computational Mathematics, Ivan Franko National University in Lviv, kom@franko.lviv.ua, http://blues.lnu.edu.ua/ami/kom/ostudin

Summary: *some typical problems in the numerical analysis of certain types of boundary value problems of the potential theory in substantially spatial formulation are considered. On the basis of the integral equation method (IE) an approximate scheme of solving one model example is built and investigated. It is also considered that the doubly connected open surface where boundary conditions are set obtains the Abelian group of symmetry of the eighth order. This article shows how using the apparatus of the group theory it is possible to solve an initial problem by the help of the sequence of the eight independent IEs, where the integration is realized only on one of the congruent constituents of the surface. It creates the conditions for two parallel processes of problem solution in general. The collocation method for obtaining approximate values of needed "density of charge distribution" in the particular two-dimensional integral equations is used. To take into account the singular way of solving the problem in the circuit of the open surface the a posteriori method of error evaluation is created and the procedure of integrating clarification of solving the task in the mesh node is implemented. To prove the reliability and estimation of the technique efficiency the number of numerical experiments is carried out including the use of so called "plane" approximation of the examined spatial problem.*

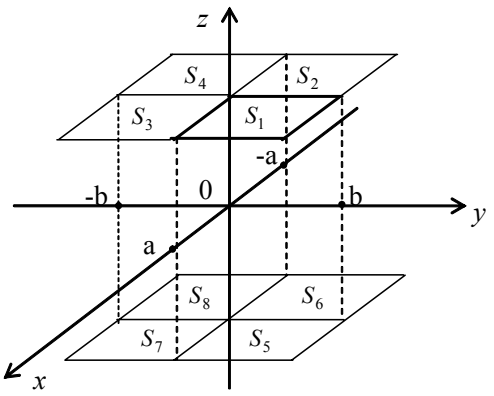
Keywords: *the potential theory, integral equations, the collocation method, the Abelian group of symmetry, matrix of Fourier transformation, the a posteriori error evaluation, integrating clarification of solving.*

INTRODUCTION

In the process of solving boundary value problems of the potential theory in electron optics the problem occurs when measuring the electrostatic field which is formed by the combination of a great number of the charged electrodes of complicated configuration. Thus, traditional application of IE method becomes substantially complicated as it is related to unknown quantities estimation mainly on the boundary surfaces. Therefore, the application of the group theory apparatus turns out to be more effective for the types of boundary tasks which obtain the Abelian group of symmetry of the finite order. However, taking into account this peculiarity in solving particular problems requires an individual approach. Although there are considerable advantages which allow to find approximate solutions with high accuracy.

1. PROBLEM STATING

Without loss of generality, we will consider the task of calculation of the electrostatic field of parallel condenser which is shown on the Pic. 1.



Pic. 1 – Position of parallel condenser disks

It is easy to see that in the mathematical modeling the information on the geometry of the charged electrodes is shown as some aggregate S of two parallel rectangular disks. These will be considered as infinitely thin disks, with two-sides

and limited piecewise smooth circuits of the finite length. From the mathematical point of view it is necessary to find a function $u(P) \in H^1(\Omega_s^-, \Delta)$, which satisfies the following conditions

$$\Delta u = 0 \quad \text{в} \quad \Omega_s^- := \mathbf{R}^3 \setminus \overline{S}; \tag{1}$$

$$\delta^{\pm} u = g_{\pm} \quad \text{on} \quad S; \tag{2}$$

$$\lim_{|P| \to \infty} u(P) = 0, \quad P \in \Omega_s^-, \tag{3}$$

where $\delta^{\pm} : H^1(\Omega_s^-) \to H^{1/2}(S)$ – the operators of track [1], $g_{\pm} \in H^{1/2}(S)$ – set functions, where $g_{\pm}$ – known value of the searching function on S from a positive and negative side accordingly, and

$$H^1(\Omega_s^-, \Delta) := \left\{ u \,\middle|\, u \in H^1(\Omega_s^-), \Delta u \in L_2(\Omega_s^-) \right\}.$$

During the electrostatic interpretation (1)-(3) $g_+(P) = g_-(P) = f(P) (P \in S)$ – the boundary potential value, which on each of two electrodes takes a permanent value. It is necessary to notice that these values in general do not possess symmetry or anti symmetry.

It is known [2] that the problem (1)-(3) is equivalent to such IE as

$$\iint_S K(P,M) \rho(M) \, \mathrm{d}S_M = f(P), \quad P \in S, \tag{4}$$

where $K(P,M) := 1/\operatorname{dist}(P,M)$ – a fundamental solution to Laplace's equation in $\mathbf{R}^3$, and $\rho(M)$ is the needed "density of charge distribution" on S.

2. ACCOUNTING PRESENT SYMMETRY

The surface S can be viewed as an aggregate of eight congruent constituents $S_i \left(i = \overline{1, 8}\right)$ (see Pic. 1). We will give (4) in accordance with a such partitioning of S as an IE system

$$\begin{cases} \sum\limits_{i=1}^{4} \iint\limits_{\Delta_i} \rho_i(x,y) K(x,y;x_0,y_0) \mathrm{d}x\mathrm{d}y + \\ + \sum\limits_{i=5}^{8} \iint\limits_{\Delta_i} \rho_i(x,y) K_h(x,y;x_0,y_0) \mathrm{d}x\mathrm{d}y = \\ = f_k(x_0,y_0), \quad (x_0,y_0) \in S_k, \; k = \overline{1, 4}; \\ \sum\limits_{i=1}^{4} \iint\limits_{\Delta_i} \rho_i(x,y) K_h(x,y;x_0,y_0) \mathrm{d}x\mathrm{d}y + \\ + \sum\limits_{i=5}^{8} \iint\limits_{\Delta_i} \rho_i(x,y) K(x,y;x_0,y_0) \mathrm{d}x\mathrm{d}y = \\ = f_k(x_0,y_0), \quad (x_0,y_0) \in S_k, \; k = \overline{5, 8}. \end{cases} \tag{5}$$

Here $\left\{\rho_i(x, y), (x, y) \in \Delta_i, \; i = \overline{1, 8}\right\}$ is a generalized "density of charge distribution" on S;

$$\Delta_1 = \Delta_5 := [0, a] \times [0, b],$$
$$\Delta_2 = \Delta_6 := [-a, 0] \times [0, b],$$
$$\Delta_3 = \Delta_7 := [0, a] \times [-b, 0],$$
$$\Delta_4 = \Delta_8 := [-a, 0] \times [-b, 0];$$

$$K(x,y;x_0,y_0) := \frac{1}{\sqrt{(x-x_0)^2 + (y-y_0)^2}};$$

$$K_h(x,y;x_0,y_0) := \frac{1}{\sqrt{(x-x_0)^2 + (y-y_0)^2 + 4h^2}};$$

$2h$ – the distance between disks.

Further, we will take into account that the surface S obtains the Abelian group of symmetry of the eighth order $\{\tau_i\}_{i=1}^8$, which, in its turn, is a direct product of Abelian sub-groups $\{e, \tau_x\}$, $\{e, \tau_y\}$, $\{e, \tau_z\}$, where e – identical transformation, and τ_x, τ_y, τ_z – mirror reflection from three pairwise orthogonal planes $\{yz\}$, $\{xz\}$, $\{xy\}$, accordingly. So that the elements of the group are the following linear transformations: $\tau_1 := e$, $\tau_2 := \tau_x$, $\tau_3 := \tau_y$, $\tau_4 := \tau_x \circ \tau_y$, $\tau_5 := \tau_z$, $\tau_6 := \tau_x \circ \tau_z$, $\tau_7 := \tau_y \circ \tau_z$, $\tau_8 := \tau_x \circ \tau_y \circ \tau_z$. As unknown functions $\rho_i(x,y)$ depend only on two independent arguments, it is obvious that

$$\tau_1 = \tau_5 = \tau_1^{-1} = \tau_5^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\tau_2 = \tau_6 = \tau_2^{-1} = \tau_6^{-1} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\tau_3 = \tau_7 = \tau_3^{-1} = \tau_7^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

$$\tau_4 = \tau_8 = \tau_4^{-1} = \tau_8^{-1} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

For further system transformation (5) in accordance with the general ideas from works [3, 4], it is reasonable to use such denotations:

$$a := \frac{1}{\sqrt{(x-x_0)^2 + (y-y_0)^2}},$$

$$b := \frac{1}{\sqrt{(x+x_0)^2 + (y-y_0)^2}},$$

$$c := \frac{1}{\sqrt{(x-x_0)^2 + (y+y_0)^2}},$$

$$d := \frac{1}{\sqrt{(x+x_0)^2 + (y+y_0)^2}},$$

$$e := \frac{1}{\sqrt{(x-x_0)^2+(y-y_0)^2+4h^2}},$$

$$f := \frac{1}{\sqrt{(x+x_0)^2+(y-y_0)^2+4h^2}},$$

$$g := \frac{1}{\sqrt{(x-x_0)^2+(y+y_0)^2+4h^2}},$$

$$h := \frac{1}{\sqrt{(x+x_0)^2+(y+y_0)^2+4h^2}}.$$

Then, in the IE system (5) we will go to the new basis:

$$\rho'_i(x,y) := \rho_i\left[\tau_i^{-1}\cdot\begin{pmatrix}x\\y\end{pmatrix}\right],\ \ i=\overline{1,8};$$

$$f'_k(x_0,y_0) := f_k\left[\tau_k^{-1}\cdot\begin{pmatrix}x_0\\y_0\end{pmatrix}\right],\ \ k=\overline{1,8}.$$

On the basis of the implemented "change of variable", which is also used in the kernels of IE, the system (5) can be shown as

$$\iint_{\Delta_1}\sum_{j=1}^{8} a'_{ij}\rho'_j(x,y)\mathrm{d}x\mathrm{d}y = f'_i(x_0,y_0),$$
$$(x_0,y_0)\in(0,a)\times(0,b),\quad i=\overline{1,8}, \tag{6}$$

where

$$A' := \left(a'_{ij}\right)_{i,j=1}^{8} = \begin{Vmatrix} a & b & c & d & e & f & g & h\\ b & a & d & c & f & e & h & g\\ c & d & a & b & g & h & e & f\\ d & c & b & a & h & g & f & e\\ e & f & g & h & a & b & c & d\\ f & e & h & g & b & a & d & c\\ g & h & e & f & c & d & a & b\\ h & g & f & e & d & c & b & a \end{Vmatrix}.$$

Thus, on this stage we have received the system of eight IEs where integration is performed only on one component surface – S_1.

In its turn, as sub-groups $\{e, \tau_x\}$, $\{e, \tau_y\}$, $\{e, \tau_z\}$ are cyclic groups, thus, the table of their characters looks the same [5]:

	e	τ
χ^1	1	1
χ^2	1	-1

,

where $\tau := \tau_x \vee \tau_y \vee \tau_z$. Thus, we can calculate the group characters of the eighth order $\{\tau_i\}_{i=1}^{8}$, taking direct product of matrices of characters (Fourier's matrices) of sub-groups $\{e, \tau_x\}$, $\{e, \tau_y\}$, $\{e, \tau_z\}$. As a result we will have

$$F := \left(F_{ji}\right)_{j,\,i=1}^{8} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1\\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1\\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1\\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1\\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1\\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1\\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1\\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \end{pmatrix}.$$

After using forward and backward transformation of Fourier, the matrix A' can be reduced to the diagonal representation [4], and the system (6) can be "split" on eight independent IEs

$$\iint_{\Delta_1} A_j\,\overline{\rho}_j(x,y)\mathrm{d}x\mathrm{d}y = \overline{f}_j(x_0,y_0),$$
$$(x_0,y_0)\in(0,a)\times(0,b),\quad j=\overline{1,8}, \tag{7}$$

where

$$\overline{\rho}_j(x,y) = \sum_{i=1}^{8} F_{ji}\cdot\rho'_i(x,y),(x,y)\in\Delta_1, \tag{8}$$

$\overline{f}_j(x_0,y_0) = \sum_{i=1}^{8} F_{ji}\cdot f'_i(x_0,y_0)$, and A_j – the elements of the diagonal matrix $F\cdot A'\cdot F^{-1}$, which are calculated after the formulas:

$$A_1 = a+b+c+d+e+f+g+h,$$
$$A_2 = a-b-c+d+e-f-g+h,$$
$$A_3 = a+b-c+d-e+f-g-h,$$
$$A_4 = a-b+c+d-e-f+g-h,$$
$$A_5 = a+b+c-d+e-f-g-h,$$
$$A_6 = a-b-c-d+e+f+g-h,$$
$$A_7 = a+b-c-d-e-f+g+h,$$
$$A_8 = a-b+c-d-e+f-g+h.$$

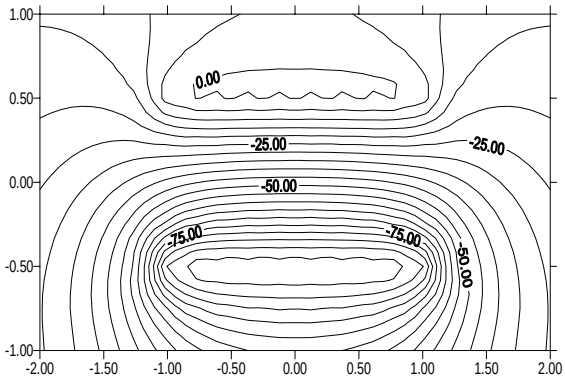
After solving approximately eight independent IEs (7) by the collocation method, and then the system (8), we will obtain the value of the functions $\rho'_i(x,y)$ $i=\overline{1,8}$, by the help of which it is possible to calculate the potential in any point of the space $(\overline{x},\overline{y},\overline{z})$ using the following formula:

$$u(\overline{x},\overline{y},\overline{z}) = \iint_{\Delta_1}\left[\frac{\rho'_1(x,y)}{\sqrt{(x'_1-\overline{x})^2+(y'_1-\overline{y})^2+(\overline{z}-h)^2}}+\right.$$
$$\left.+\frac{\rho'_2(x,y)}{\sqrt{(x'_2+\overline{x})^2+(y'_2-\overline{y})^2+(\overline{z}-h)^2}}+\right.$$

$$+\frac{\rho'_3(x,y)}{\sqrt{(x'_3-\bar{x})^2+(y'_3+\bar{y})^2+(\bar{z}-h)^2}}+$$
$$+\frac{\rho'_4(x,y)}{\sqrt{(x'_4+\bar{x})^2+(y'_4+\bar{y})^2+(\bar{z}-h)^2}}+$$
$$+\frac{\rho'_5(x,y)}{\sqrt{(x'_5-\bar{x})^2+(y'_5-\bar{y})^2+(\bar{z}+h)^2}}+$$
$$+\frac{\rho'_6(x,y)}{\sqrt{(x'_6+\bar{x})^2+(y'_6-\bar{y})^2+(\bar{z}+h)^2}}+$$
$$+\frac{\rho'_7(x,y)}{\sqrt{(x'_7-\bar{x})^2+(y'_7+\bar{y})^2+(\bar{z}+h)^2}}+$$
$$+\frac{\rho'_8(x,y)}{\sqrt{(x'_8+\bar{x})^2+(y'_8+\bar{y})^2+(\bar{z}+h)^2}}\Bigg]\mathrm{d}x\mathrm{d}y,$$

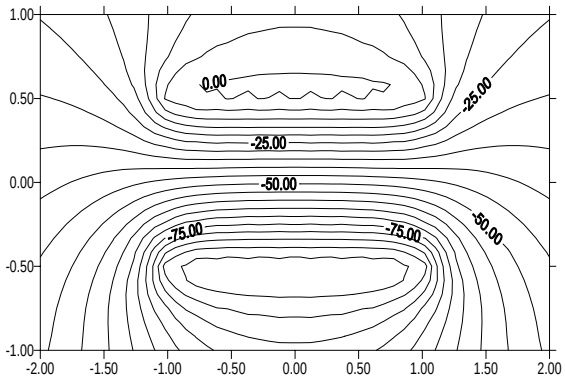
where $\begin{pmatrix} x'_i \\ y'_i \end{pmatrix} = \tau_i^{-1}\begin{pmatrix} x \\ y \end{pmatrix}$, $i=\overline{1,8}$.

The use of the approach offered above allows to test an initial task for the arbitrary boundary values of the potential. In the Pic. 2 – 3 the distributing of lines of even potential is presented in the plane $x=0$ at different relations of the length (a) and width (b) of the plate, with the use of piecewise permanent approximation of IE density with the proper boundary values of the potential: $f_1=1$; $f_2=-100$. A number of collocation points at the separation of constituent S_1 equals 100.

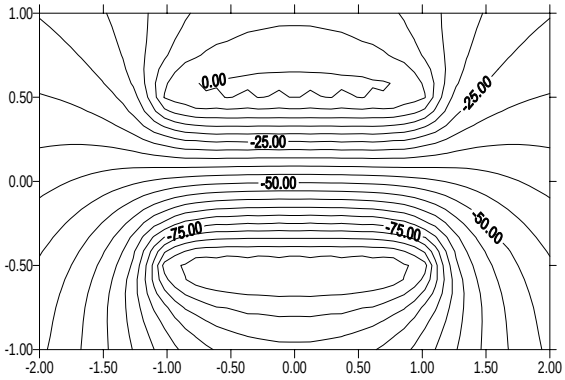


Pic. 2 – Distributing of level lines in the plane $x=0$ at $a/b=1$

According to [6], for showing of the qualitative picture of the field in the plane, which forms as a result of central cross-section of the investigated system of the charged electrodes it is appropriate to use some "plane" approximation when solving a spatial problem. The Pic. 4 shows the distribution of lines of even potential when solving the proper "plane" task, and the Table 1 shows the value of the potential in some checkpoints when solving a spatial problem and proper "plane" approximation.



Pic. 3 – Distributing of level lines in the plane $x=0$ at $a/b=16$



Pic. 4 – Distributing of level lines of "plane" task

Table 1. Potential of the electrostatic field in the points of plane $x=0$

(y, z)	u $(a/b=16)$	"plane" approximation
(-1.00, -0.75)	-79.3731	-79.3729
(-1.00, -0.25)	-66.4593	-66.4594
(-1.00, 0.00)	-46.3909	-46.3910
(-1.00, 0.25)	-27.0001	-27.0004
(-1.00, 0.75)	-10.4086	-10.4082
(-0.75, -0.50)	-99.9925	-99.9930
(1.00, 0.75)	-10.4086	-10.4082
(2.00, -1.00)	-48.9304	-48.9307

It is necessary to notice that solving a spatial problem with the arbitrary boundary values of potentials on electrodes we should calculate a special additive constant C which appears in the integral presentation of the field [6]. In our case $C=-49.5$.

3. A POSTERIORI METHOD OF ERROR EVALUATION

Without diminishing general attitude, we will consider one of the eight independent IEs (7):

$$\left(K\bar{\rho}_1\right)(x_0,\, y_0) \equiv \iint\limits_{\Delta_1} A_1 \bar{\rho}_1(x,y)\,\mathrm{d}x\mathrm{d}y = \\ = \bar{f}_1(x_0, y_0), \quad (x_0, y_0) \in (0,\, a)\times(0,\, b). \tag{9}$$

An approximate solution to this equation is found by the collocation method with the use of piecewise permanent basic functions. Element division is conducted according to $\Delta_1 = [0,\, a]\times[0,\, b]$. By "Extremal" is considered an element $D_{N_xN_y} := [a - h_x,\, a]\times[b - h_y, b]$, which corresponds the angular point of constituent S_1 (see. Pic. 5). Here $h_x = \dfrac{a}{N_x}$, $h_x = \dfrac{b}{N_x}$, where N_x, N_y – the number of points of segment division $[0,\, a]$, $[0,\, b]$, accordingly. It is necessary to notice that without accounting the present symmetry in geometry of the open surfaces we should have taken into account the peculiarities of solving in the circuit of eight angular points of surface S, which substantially complicates all algorithm of calculations.

It is known [7] that the "density of charge distribution" $\bar{\rho}_1(x,y)$ in the circuit of the "angular point" $Q := (a,b)$ and reaching the circuit has a feature which is expressed by the formula

$$\Omega(x,\, y) = \frac{[(a-x)\cdot(b-y)]^{1/2}}{(a-x)^{\gamma} + (b-y)^{\gamma}},$$

where $\gamma \approx 1{,}7034$. To remove this peculiarity it is necessary to introduce a special change of variables in the proper double integrals. However, it considerably complicates the algorithm of approximate solving of the integral equation and does not enable to use a simple and effective numeral and analytical scheme [8]. Therefore, we will take into account this peculiarity by net condensing in the circuit of the point Q to achieve the accuracy of calculations and we will use general ideas mentioned in the works [9, 10].

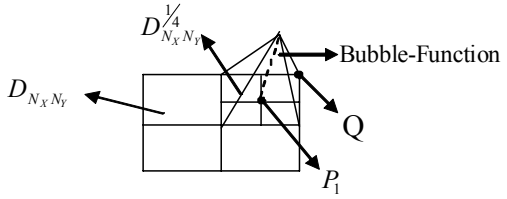
Let us suppose that in the result of solving the systems of linear algebraic equations, which approximates the proper operator equation, we get an approximate solution $\bar{\rho}_{1\varepsilon}(P)$, the analysis of which in any case requires the investigation of properties of its error $\varepsilon_{\rho_1} := \bar{\rho}_1 - \bar{\rho}_{1\varepsilon}$.

Let us put some limits and create approximation e_{ρ_1} to the real error ε_{ρ_1}. The verification of satisfying boundary condition consists in calculating the potential in some checkpoint P_1, which lies in the circuit of the angular point of the constituent S_1 and corresponds to the "extremal" element $D_{N_xN_y}$. Then we will give the error function e_{ρ_1} as $e_{\rho_1} = \lambda_1 \cdot B_{P_1}(x,\, y)$ and we will find an unknown parameter λ_1 by collocation (9) exactly in the point $P_1 := \left\{a - \dfrac{h_x}{4},\, b - \dfrac{h_y}{4}\right\}$:

$$\lambda_1 = \frac{\bar{f}_1(P_1) - \left(K\bar{\rho}_1\right)(P_1)}{\left(K B_{P_1}\right)(P_1)}. \tag{10}$$

It should be noticed that such simple formula for parameter λ_1 calculating is connected with finiteness of the bilinear bubble-function $B_{P_1}(x,\, y)$:

$$\operatorname{supp} B_{P_1}(x,y) = D^{1/4}_{N_xN_y} := \\ = \left[a - \frac{h_x}{2},\, a\right]\times\left[b - \frac{h_y}{2},\, b\right].$$



Pic. 5 – Schematic presentation of the "extremal" element

Calculating the denominator in the formula (10) it is convenient to take an element $D^{1/4}_{N_xN_y}$ to the local coordinate system (α, β) so that $|\alpha| \le 1$, $|\beta| \le 1$. It enables the presentation of $\left(K B_{P_1}\right)(P_1)$ as a sum of four integrals I_1, I_2, I_3, I_4:

$$\mathrm{I}_1 := \int\limits_{x_i - \tilde{\Delta}_1}^{x_i} \int\limits_{y_j - \tilde{\Delta}_2}^{y_j} A_1 \cdot B^1_{P_1}(x,\, y)\,\mathrm{d}x\mathrm{d}y,$$

$$\mathrm{I}_2 := \int_{x_i}^{x_i+\widetilde{\Delta}_1} \int_{y_j-\widetilde{\Delta}_2}^{y_j} A_1 \cdot B_{P_1}^2(x,\, y)\mathrm{d}x\mathrm{d}y,$$

$$\mathrm{I}_3 := \int_{x_i}^{x_i+\widetilde{\Delta}_1} \int_{y_j}^{y_j+\widetilde{\Delta}_2} A_1 \cdot B_{P_1}^3(x,\, y)\mathrm{d}x\mathrm{d}y,$$

$$\mathrm{I}_4 := \int_{x_i-\widetilde{\Delta}_1}^{x_i} \int_{y_j}^{y_j+\widetilde{\Delta}_2} A_1 \cdot B_{P_1}(x,\, y)\mathrm{d}x\mathrm{d}y,$$

where $\widetilde{\Delta}_1 = \frac{\mathrm{h_x}}{4}$, $\widetilde{\Delta}_2 = \frac{\mathrm{h_y}}{4}$, in each of them to perform some change of variables. So, for example, for I_1 we will have

$$\begin{cases} x(\alpha,\, \beta) := \dfrac{2x_i - \widetilde{\Delta}_1}{2} + \dfrac{\widetilde{\Delta}_1}{2} \cdot \alpha, \\ y(\alpha,\, \beta) := \dfrac{2y_j - \widetilde{\Delta}_2}{2} + \dfrac{\widetilde{\Delta}_2}{2} \cdot \beta, -1 \le \alpha, \beta \le 1. \end{cases}$$

Thus, the proper constituents of the bubble-function in the local coordinates are as following

$$B^1(\alpha,\, \beta) = \frac{1}{4}(1+\alpha)(1+\beta),$$

$$B^2(\alpha,\, \beta) = \frac{1}{4}(1-\alpha)(1+\beta),$$

$$B^3(\alpha,\, \beta) = \frac{1}{4}(1-\alpha)(1-\beta),$$

$$B^4(\alpha,\, \beta) = \frac{1}{4}(1+\alpha)(1-\beta).$$

We use the above mentioned values for a posteriori error evaluation on the "extremal" element to find the approximate solution (9) with the beforehand accuracy. As a criterion of stopping the clarification process of the approximate solution $\overline{\rho}_{1\varepsilon}(P)$ there is some indicator

$$\eta_{\rho_1} := \frac{\left\| e_{\rho_1}(x,y) \right\|_{L_2}}{\sqrt{\left\| e_{\rho_1}(x,y) \right\|_{L_2}^2 + \left\| \overline{\rho}_{1\varepsilon}(x,y) \right\|_{L_2}^2}} \cdot 100\%. \quad (11)$$

If the indicator (11) does not exceed set possible level, then we stop the clarification $\overline{\rho}_{1\varepsilon}$, otherwise we carry out the net condensing. In our case, on the first step of net condensing we will get four new elements $\left(D^{1/4}_{N_x N_y} \right)_i$, $(i = \overline{1,4})$.

The Table 2 shows the value of the potential in some checkpoints which are situated in the circuit of the "extremal" element before (u_0) and after (u_1) using a posteriori error appraiser at a possible level of 10 %.

Table 2. Potential at some checkpoints in the case of piecewise-bilinear bubble function

N_y	y	0,9375	0,8750	0,8125
	z	0,9375	0,9375	0,9375
4	u_0	0,8963	0,9454	0,9697
	u_1	1,0358	1,0189	1,0105
	л$_1$	0,1395	0,0735	0,0408
	з	0,2116	0,1132	0,0631
5	u_0	0,9321	0,9792	–
	u_1	1,0234	1,0072	–
	л$_1$	0,0913	0,0280	–
	з	0,1365	0,0421	–
6	u_0	0,9600	–	–
	u_1	1,0138	–	–
	л$_1$	0,0538	–	–
	з	0,0794	–	–

4. ITERATION REFINEMENT OF SOLUTIONS

Using the method of a posteriori evaluation of error partly offered above, we will consider the iteration process of clarification of the approximate solutions to the problem (9). On the basis of preliminary received densities using the collocation method $\overline{\rho}_j(x,y)$ $j = \overline{1,\, 8}$ з (7) in the points $(x_i,\, y_p)$, where

$$x_i = (2i-1)\cdot\frac{h_x}{2},\ i = \overline{1,\, N_x},$$

$$y_p = (2p-1)\cdot\frac{h_y}{2},\ p = \overline{1,\, N_y},$$

we find the value $\overline{\rho}_j(x,y)$ in the mesh nodes $(\widetilde{x}_k,\, \widetilde{y}_l)$, where $\widetilde{x}_k = (k-1)\cdot h_x$, $k = \overline{1,\, N_x+1}$, $\widetilde{y}_l = (l-1)\cdot h_y$, $l = \overline{1,\, N_y+1}$, as follows:

1) in the extreme points of the net, for example, in the point $(\widetilde{x}_1,\, \widetilde{y}_{N_y+1})$, $\overline{\rho}_j(\widetilde{x}_1,\, \widetilde{y}_{N_y+1}) := \overline{\rho}_j(x_1,\, y_{N_y})$;

2) for the points $(\widetilde{x}_k,\, \widetilde{y}_{N_y+1})$, $k = \overline{1,\, N_x}$,

$$\overline{\rho}_j(\widetilde{x}_k,\, \widetilde{y}_{N_y+1}) := \frac{1}{2}\left[\overline{\rho}_j(x_i,\, y_{N_y}) + \overline{\rho}_j(x_{i+1},\, y_{N_y})\right],$$

where $i = \overline{1,\, N_x-1}$;

3) in the points $(\widetilde{x}_k,\, \widetilde{y}_l)$, $k = \overline{2,\, N_x}$, $l = \overline{2,\, N_y}$, we perform

$$\overline{\rho}_j(\widetilde{x}_k,\, \widetilde{y}_l) := \frac{1}{4}\left[\overline{\rho}_j(x_i,\, y_p) + \overline{\rho}_j(x_{i+1},\, y_p) + \overline{\rho}_j(x_i,\, y_{p+1}) + \overline{\rho}_j(x_{i+1},\, y_{p+1})\right],$$

where $i = \overline{1,\, N_x-1}$, $p = \overline{1,\, N_y-1}$.

Then, we calculate the potential in the mesh

nodes $(\tilde{x}_k, \tilde{y}_l)$. On the basis of the bubble-function, by analogy with (10), we receive $\tilde{\lambda}_j$. On the basis of the discovered values of density in the mesh nodes we re-count $\overline{\rho}_j(x,y)$ in the points (x_i, y_p) after the rules described above 1)-3). Then the procedure of iteration clarification can repeat. The criterion of stopping of such calculations can be, for example, fixing of a number of iterations or under the condition of $\left\| e_{\rho_j}(x,y) \right\|_{L_2} \le eps$.

The Table 3 shows the potential value in some checkpoints before and after the use of iteration refinement of solutions.

Table 3. Potential *u* at some checkpoints after the use of iteration refinement of solution (eps = 0,01)

y	*z*	Iteration			
		1	2	3	4
0,95	0,95	0,9029	0,8150	0,9357	0,9851
0,80	0,95	0,9650	0,8851	1,0318	–
0,60	0,95	0,9545	0,8892	1,0689	–
0,40	0,95	0,9569	0,8868	1,0786	–
0,20	0,95	0,9579	0,8876	1,0851	–
0,80	0,80	1,0338	–	–	–
0,60	0,60	1,0000	–	–	–

5. CONCLUSIONS

Taking an example of the numerical solving one spatial problem the problems at calculating of the electrostatic field are investigated. Approximate solutions are received on the basis of the IE method taking into account the present symmetry in geometry of separate elements of the surface. It allowed to carry on solving the eight independent IEs, where integration is conducted for $1/8$ of the whole surface. It, in its turn, enables to decrease random-access memory of the computer when forming the system of linear algebraic equations, which approximates the proper integral equation in 64 times and to avoid numerical instability which can arise at slight increase of the solved systems. This approach also enables instead of eight special points of the surface where it is necessary to take into account the singular way of the solution to control only one. In the circuit of this "extremal" element a feature is taken into account on the basis of the created a posteriori method of error evaluation. For clarification of the solutions in the mesh nodes the iteration process is introduced. All the advantages of this method were confirmed by numerical experiments.

6. REFERENCES

[1] *Lions J.-L., Magenes E.* Problйmes aux Limites non Homogйnes et Applications. Vol. 1. Dunod. Paris, 1968.

[2] *Sybil Yu.M.* Three dimensional elliptic boundary value problems for an open Lipschitz surface // Mathematical Studies. 1997. Vol. 8. № 2. P. 79−96.

[3] *Zakharov Y.V., Safronov S.I., Tarasov R.P.* The Method of numeral solving of integral equations in the boundary problems with the Abelian group of symmetries of the finite order // Journal of Computational Mathematics and Mathematical Physics. 1990. Vol. 30. № 11. P. 1661–1674 (in Russian).

[4] *Zakharov Y.V., Safronov S.I., Tarasov R.P.* The Abelian groups of the finite order in the numerical analysis of linear boundary problems of the potential theory // Journal of Computational Mathematics and Mathematical Physics. 1992. Vol. 32. № 1. P. 40–58 (in Russian).

[5] *Serr G.P.* Linear presentations of finite groups. M.: RKHD, 2003. 132 p. (in Russian).

[6] *Harasym Y.S., Mochurad L.I., Ostudin B.A.* Paralling of the method of integral equations solving the boundary problems of the potential theory of potential with the symmetric boundary elements // Materials of the Third International Scientific and Technical Conference "Computer Sciences and Information Technologies", September, 25-27 2008, Lviv: "Tower and Co". – 2008. – P. 342-346 (in Ukrainian).

[7] *Morrison J.A., Lewis J.A.* Charge singularity at the corner of a fllat plate // SIAM J. Appl. Math. Vol. 31. № 2. 1976. P. 233-249.

[8] *Garasym Ya.S., Ostudin B.A.* On numerical appoach to solve some three-dimensional boundary value problems in potential theory based on integral equation method // Journal of Computational and Applied Mathematics. 2003. № 1 (88). P. 17−28.

[9] *Eriksson K., Estep D., Hansbo P., Johnson C.* Introduction to Adaptive Method for Differential Equation // Acta Numerica. – 1955. P. 1-54.

[10] *Carstensen C., Maischak M., Praetorins D., Stephan E.P.* Residual based a posteriori estimate for hypersingular equation on surfaces // Numer. Math. – 2004. – 97. P. 397-425.

Lesya Mochurad *was born in Novyy Rozdil, Lviv Region, Ukraine in April 14, 1984. In 2006 she graduated from Ivan Franko National University in Lviv, Faculty of Applied Mathematics and Computer Science (master's degree). Since 2006 she has been a post-graduate student at Ivan Franko National University in Lviv, Faculty of Applied Mathematics and Computer Science. Scientific interests: method of the numerical solving integral equations in the boundary problems of the potential theory with the Abelian group of symmetries of the finite order and decomposition of complicated areas.*

Yaroslav Harasym *was born in Stavropol Territory, the Russian Federation, in March 29, 1969, Higher Education. Since 2008 he has been working as a senior teacher at the department of computational mathematics at Ivan Franko National University in Lviv. Scientific interests: numerical solving integral equations in the boundary problems in mathematical physics for open surfaces.*

Borys Ostudin *was born in Lviv May 3, 1947, Candidate of Physical and Mathematical Sciences degree. Since 1982 he has been working as an associate professor at the department of computational mathematics at Ivan Franko National University in Lviv. Scientific interests: numerical solving one- and two-dimensional integral first kind equations with weak singularities in kernels.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

ДИСТРИБУТИВНА СЕНСОРНА МЕРЕЖА ДЛЯ СИСТЕМ БЕЗПЕКИ

Павло Биковий

Науково-дослідний інститут інтелектуальних комп'ютерних систем
Тернопільський національний економічний університет
pb@tneu.edu.ua

Резюме: *розроблено дешевий мережевий контролер для сповіщувачів систем безпеки. Особливістю даного контролера є підтримка функціонування в двопровідній мережі топології "спільна шина", що значно зменшує кількість ліній передачі даних від сповіщувачів.*

Ключові слова: *система безпеки, мережевий контролер, двопровідна мережа.*

1. ВСТУП

На сьогодні системам безпеки і захисту від несанкціонованого доступу приділяється значна увага [1]. Крім стрімкого росту кількості таких систем, їх функціональні можливості суттєво розширюються [2], що забезпечує комплексний захист від усіх видів загроз на всіх можливих шляхах їх виникнення. Очевидно, що при цьому системи безпеки можуть значно ускладнюватись та зростати в ціні, тому створюються системи їх автоматизованого проектування [3, 4] з подальшим виявленням оптимальних по ціні, якості і надійності.

Однією з основних умов масового використання систем безпеки і захисту від несанкціонованого доступу є їх максимальна уніфікація і стандартизація. Основна обробка сигналу, щодо виявлення загрози, зазвичай проводиться в сповіщувачі (сенсорному вузлі) [5]. Такий сповіщувач є досить складним пристроєм, в структурі якого можна виділити чотири функціональні вузли [5, 6]: (i) один або декілька чутливих елементів (сенсорів), які сприймають інформацію про загрозу і перетворюють її в електричний сигнал (як правило, за назвою цих чутливих елементів називають сповіщувач); (ii) схему аналогової обробки вихідного електричного сигналу чутливого елемента, яка виконує функції підсилення, фільтрування, селекції та процесор, що приймає рішення по виявленню наявності загрози; (iii) уніфіковану вихідну схему сповіщувача, що формує вихідні сигнали сповіщувача та представляє собою нормально замкнуті або розімкнуті контакти, які змінюють стан при виявлені загрози; (iv) схему захисту від несанкціонованого доступу в сповіщувач і порушення його функціонування.

Сповіщувачі з допомогою ліній зв'язку підключаються до приймально-контрольного приладу (централі). Кожному сповіщувачу під'єднують стільки двопровідних ліній зв'язку, скільки чутливих елементів він має, а також додаткову двопровідну лінію живлення. Наприклад, пасивні інфрачервоні сповіщувачі типу SRPG-2N фірми CROW [7] мають інфрачервоний чутливий елемент, мікрофонний чутливий елемент, а також мікроперемикач захисту від відкриття корпусу сповіщувача і вимагають для підключення чотирьох двопровідних ліній зв'язку (з врахуванням додаткової допровідної лінії живлення).

Для постійного контролю стану ліній зв'язку і захисту їх від пошкодження порушником безпеки вхідні каскади приймально-контрольного приладу налаштовані на фіксоване значення опору зовнішньої лінії. Переважна більшість виробників випускає приймально-контрольні прилади налаштовані на значення опору 2 кОм.

Зрозуміло, що розглянуті комплексні системи безпеки мають мати значну кількість каналів для диференціації загроз на окремих напрямах і ділянках. Традиційно системи безпеки будуються по топології "зірка", в яких сумарна довжина ліній зв'язку відповідає сумі відстаней до всіх сповіщувачів[1, 2]. Тому затрати на кабель зв'язку становлять значний відсоток ціни системи безпеки. Особливо це проявляється в системах безпеки периметру або територій з великою кількістю приміщень. Для зменшення кількості ліній зв'язку можливе послідовне включення

вихідних схем сповіщувачів послідовно. Тоді спрацювання чутливого елемента довільного сповіщувача приведе до ідентифікації загрози приймально-контрольним приладом. Однак, розпізнати конкретний напрям і вид загрози не вдається – система стає дешевшою і менш інформативною, і тому виникає протиріччя між затратами на лінії зв'язку і інформаційною здатністю системи.

Вирішення цього протиріччя можливе за рахунок мережевих технологій з використанням провідного каналу зв'язку. Наявність на ринку великої номенклатури різноманітних мікроконтролерів дозволяє обладнати ними сповіщувач і створювати системи безпеки, які базуються на топології "спільна шина". Тоді сумарна довжина ліній зв'язку буде визначатися відстанню між найбільш віддаленими між собою сповіщувачами та сумою довжин відгалужень до відповідних сповіщувачів. Наявність гальванічного зв'язку між приймально-контрольним приладом і сповіщувачем дозволяє організувати живлення сповіщувача через провідники мережі.

Таким чином, метою статті є розробка архітектури дистрибутивної сенсорної мережі (ДСМ) на основі сповіщувачів систем безпеки, яка описана нижче.

2. АРХІТЕКТУРА РОЗРОБЛЕНОЇ ДИСТРИБУТИВНОЇ СЕНСОРНОЇ МЕРЕЖІ

В склад даної архітектури розробленої ДСМ доцільно включити запропонований мережевий контролер та адаптер мережі [8]. Мережевий контролер повинен забезпечити наступні функції:

- проводити обробку вихідних сигналів сповіщувачів безпеки різного типу і принципу дії;
- підтримувати роботу в спеціалізованій провідній мережі сповіщувачів безпеки з топологією "спільна шина" та живитись від цієї ж мережі.

Адаптер мережі повинен забезпечити обмін даними та живлення контролерів і сповіщувачів мережі.

2.1 УЗАГАЛЬНЕНА СТРУКТУРА СТАНДАРТНОГО КОМПОНЕНТА РОЗРОБЛЕНОЇ ДИСТРИБУТИВНОЇ СЕНСОРНОЇ МЕРЕЖІ

Узагальнена структура стандартного компонента розробленої дистрибутивної сенсорної мережі на базі запропонованого мережевого контролера (UNC), інтегрованого в склад сповіщувача (сенсорного вузла, SU) включає (рис. 1) чутливі елементи S1…Sn сповіщувачіа SU, з'єднані із схемою аналогової обробки та процесором (APCP) яка передає результат (виявлення/не виявлення) вихідній схемі сповіщувача (OCSU). Мікроконтролер MC опитує виходи OCSU та взаємодіє з мережею використовуючи апаратні засоби послідовного двопровідного інтерфейсу IF HW. Джерело живлення PS забезпечує живлення усіх елементів.

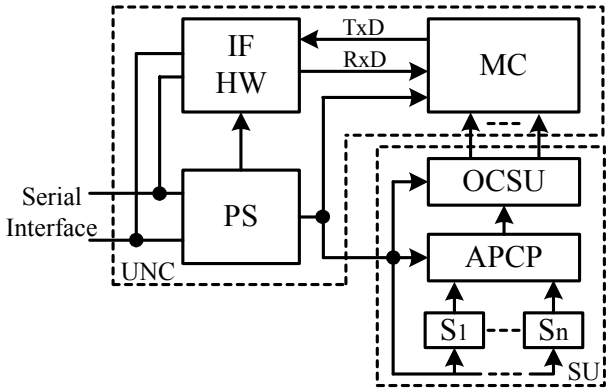


Рис. 1 – Архітектура стандартного компонента розробленої ДСМ

Одним з основних питань, які необхідно вирішити при побудові ДСМ, є питання вибору типу інтерфейсу. Слід відзначити, що основна характеристика інтерфейсу – пропускна здатність, в даному випадку не може бути критерієм відбору, оскільки потрібна для систем безпеки пропускна здатність є малою. Кількість сповіщувачів в системі безпеки зазвичай не перевищує сотні, кожен сповіщувач має не більше чотирьох чутливих елементів, а період опитування сповіщувачів - одна секунда, є цілком прийнятним. В даному випадку важливішими є вимоги топології "спільна шина": (i) достатня потужність сигналу для живлення сповіщувачів (навантажувальна здатність); (ii) простота і низька вартість адаптації до вимог замовника.

Аналіз відомих типів послідовних інтерфейсів показав, що мінімальні затрати обладнання забезпечує інтерфейс 1-Wire фірми Dallas Semiconductors [9]. Однак цей інтерфейс призначено для комунікації на відносно малі відстані. Тому в розробленому контролері доцільно використати модифікований послідовний асинхронний інтерфейс RS-232C [10, 11]. В той же час живлення контролера доцільно організувати аналогічно до інтерфейсу 1-Wire. Таке рішення має наступні переваги: (i) широке використання інтерфейсу RS-232C у комп'ютерному обладнанні; (ii) легкість забезпечення підтримки інтерфейсу RS-232C на

програмному рівні; (iii) можливість встановлення частоти опитування сповіщувачів та швидкості обміну даними згідно потреб користувача, що дозволить використати довільні типи кабелю; (iv) велика амплітуда імпульсів (±12 В) і можливість формування потужних вихідних імпульсів адаптера (сотні мА), що дозволить реалізувати живлення контролерів від мережі; (v) мала складність апаратного забезпечення.

В процесі функціонування мережі сервер почергово опитує мережеві контролери системи безпеки, посилаючи відповідний запит за допомогою послідовного інтерфейсу. Запит сервера через апаратні засоби провідного інтерфейсу IF HW поступає на мікроконтролер MC, який опитує виходи вихідної схеми сповіщувача на спрацювання чутливих елементів S1…Sn і формує байт відповіді. Одночасно схема живлення PS формує з вхідного повідомлення (запиту сервера) живлення для MC та сповіщувача.

Всі операції, які виконує MC, не виходять за рамки традиційних завдань мікроконтролерів. Оригінальними вузлами в розробленому універсальному контролері є IF HW і PS, які повинні забезпечити функціонування контролера при живленні від мережі, а також адаптер мережі, який підключається до сервера. Розглянемо побудову цих вузлів детальніше.

2.2 СХЕМА МЕРЕЖЕВОГО КОНТРОЛЕРА

Джерело живлення (PS, див. рис. 1) у складі принципової схеми розробленого мережевого контролера(рис. 2) включає вхідну розв'язку V1, C1 і стабілізатор живлення VR з вихідним конденсатором C2. Логічні нулі, які приходять по мережі на базі модифікованого послідовного інтерфейсу RS-232C, мають амплітуду +12 В і заряджають конденсатор C1. Останній підтримує напругу на вході стабілізатора достатньою для того, щоб на його виході формувалася напруга +5 В, необхідна для живлення мікроконтролера MC і сповіщувача безпеки.

Апаратні засоби послідовного інтерфейсу IF HW включають:

- приймач на транзисторі V2 і резисторах R1, R2, який представляє собою інвертор, рівень вихідного сигналу якого узгоджений по напрузі з допустимим на вході RxD мікроконтролера MC;
- передавач на транзисторі V4, стабілітроні V3 і резисторах R3, R4, який представляє собою підсилювач-інвертор вихідних сигналів мікроконтролера MC (0 В і +5 В) до рівня, який відповідає сигналам мережі (+12 В і –12 В). Надійне запирання транзистора V4 досягається наявністю стабілітрона V3 з напругою стабілізації, близькою до 8 В.

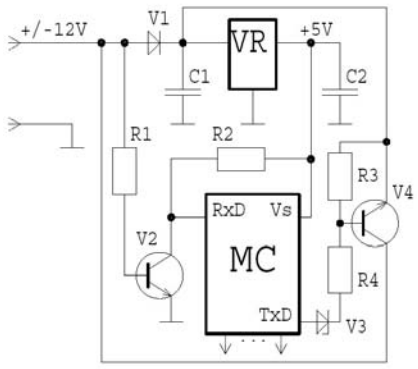


Рис. 2 – Принципова схема розробленого мережевого контролера

Важливою перевагою запропонованої схеми блоку живлення і апаратних засобів інтерфейсу є автоматичне відслідковування амплітуди імпульсів при відповіді контролера за амплітудою вхідних імпульсів, що заряджають конденсатор C1. Різниця амплітуд, яка відповідає спаду напруги на діоді V1 і транзисторі V4, не дозволяє вихідним імпульсам інтерфейсу контролера заряджати конденсатори C1 інших контролерів. В той же час, імпульси +12 В звертання сервера підзаряджають конденсатори C1 всіх контролерів в мережі незалежно від того, до якого контролера сервер звертається з запитом. Тому в мережі відсутній перерозподіл заряду і всі контролери з точки зору живлення працюють індивідуально.

2.3 СХЕМА МЕРЕЖЕВОГО АДАПТЕРА

Важливу роль в розробленій мережі на базі модифікованого послідовного інтерфейсу RS-232C відіграє адаптер мережі, який підключається до сервера. Принципова схема розробленого мережевого адаптера(рис. 3) включає:

- блок живлення, що складається з трансформатора мережі T1, двополярного випрямляча, реалізованого на діодах V1, V2 і конденсаторах фільтрів C1, C2;
- двокаскадний підсилювач, реалізований на транзисторах V3, V5 і резисторах R1…R3. Він служить для формування потужних вихідних сигналів інтерфейсу RS-232C для заряду конденсаторів C1 мережевого контролера (див. рис. 2);
- стабілізатор струму розряду лінії (встановлення логічної 1), реалізований на польовому транзисторі V6 і резисторі R5;
- схему захисту від коротких замикань,

реалізовану на транзисторі V4 і резисторі R4;
- схему захисту від перенапружень, реалізовану на симетричному стабілітроні V7 і резисторі R6.

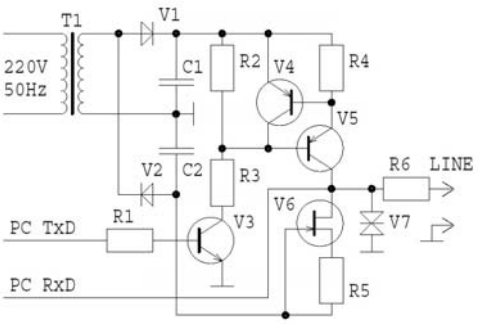


Рис. 3 – Принципова схема адаптера мережі

3. РЕАЛІЗАЦІЯ ДСМ НА БАЗІ СПОВІЩУВАЧІВ БЕЗПЕКИ

Для оптимізації використання потужності живлення в розроблених контролерах вжиті наступні заходи:
- в мережевому контролері використано мікроконтролер типу ATtiny2313 фірми Atmel [12], в якому оптимально поєднуються необхідні функціональні можливості, малий струм споживання і низька ціна;
- частота кварцового резонатора контролера вибрана пониженою – 4 МГц, що також зменшує струм споживання мікроконтролера;
- передача запитів сервера здійснюється при швидкості інтерфейсу 1200 Бод, а передача повідомлень розробленого мережевого контролера - при швидкості інтерфейсу 9600 Бод. Це дозволяє збільшити відносний час заряду конденсатора C1 блока живлення контролера (див. рис. 2);
- інтерфейс RS-232C починає повідомлення стартовим імпульсом (логічний нуль, напруга якого відповідає +12 В), далі ідуть імпульси, що відповідають молодшим розрядам байта, який передається. Тому, для покращення заряду конденсатора C1 блоку живлення (див. рис. 2), програмним шляхом перші два розряди встановлюють в нуль. Це обмежує кількість контролерів під'єднаних до одного адаптера до 64. Однак така кількість цілком прийнятна. Збільшення їх кількості не можливе через зростання енергоспоживання, що не дозволять розмістити на одній двопровідній лінії більше контролерів;
- струм розряду лінії (визначається опором резистора R5) вибраний мінімальним для даної лінії і заданої швидкодії інтерфейсу – 1,5 мА.

Стандартний компонент розробленої ДСМ (рис.1), що включає сповіщувач безпеки SRPG-2N та мережевий контролер (рис. 4), може працювати в двопровідних мережах з топологією "спільна шина", що забезпечує живлення сповіщувачів через мережу. Загальна кількість сповіщувачів не перевищує 64 штук, а кожен сповіщувач може мати декілька чутливих елементів.

Рис. 4 – Компонент ДСМ та базі сповіщувача безпеки SRPG-2N та розробленого мережевого контролера

Структурна схема ДСМ (рис.5) включає комп'ютер (сервер мережі), що виконує функції приймально-контрольного приладу, оснащений одним (або декількома) COM-портами до якого підключається адаптер мережі, котрий в свою чергу з'єднаний двопровідною лінією з кожним компонентом ДСМ.

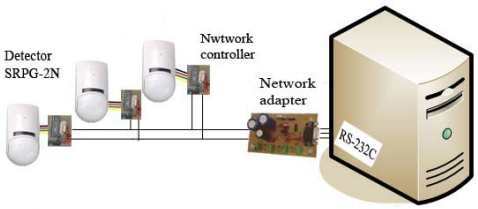


Рис. 5 – Структурна схема розробленої ДСМ

3.1 ОЦІНКА ЕФЕКТИВНОСТІ ПРОПОНОВАНОГО РІШЕННЯ

Оцінку ефективності пропонованого рішення ДСМ виконаємо на прикладі об'єкту (рис. 6), що налічує 18 кімнат. При цьому необхідно знайти сумарну довжину кабелю топології "зірка" і порівняти її із сумарною довжиною запропонованої топології "спільна шина". Для визначення довжини кабелю застосовують метод сумування та статистичний метод [13]. Статистичний метод характеризується меншою точністю, але він не є трудомістким і зазвичай використовується для великих об'єктів. В нашому випадку (нескладного об'єкту) доцільно застосувати більш точний метод сумування, згідно якого сумарна довжина кожного окремого кабелю відповідно становить: 183,3 м для топології "зірка", і 108,6 м для топології "спільна

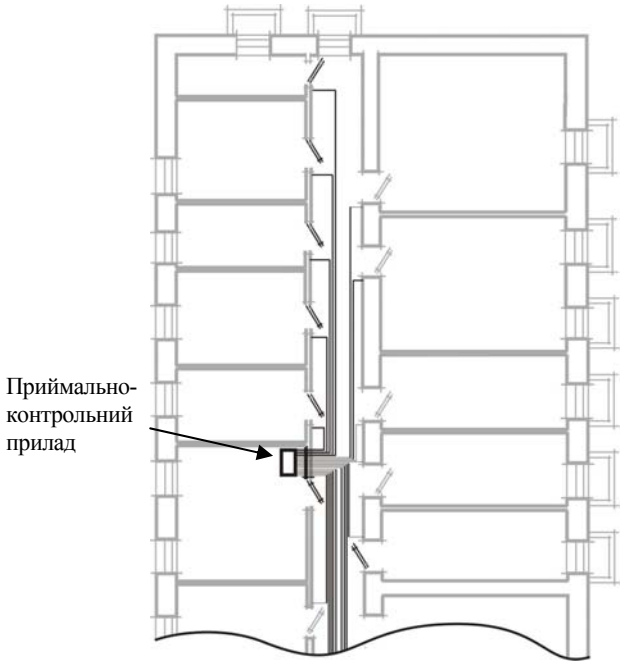


Рис. 6 – Приклад розміщення кімнат на об'єкті та кабелів по топології "зірка".

шина". Із проведеного розрахунку видно, що застосування мережевого контролера, який забезпечує підтримку роботи в двопровідній мережі сповіщувачів безпеки з топологією "спільна шина", дозволяє зменшити дожину кабелю в 1,7 раза (для даного випадку). Слід зазначити, що чим більша відстань від приймально-контрольного приладу до сповіщувачів безпеки, тим доцільніше використання мережевого контролера. Крім того, застосування мережевого контролера дає можливість зменшити витрати на приймально-контрольний прилад (його функції виконує комп'ютер), акумулятор з блоками розширення і рідкокристалічне табло з клавіатурою.

ПОДЯКИ

Робота виконується за підтримки Міністерства освіти і науки України та Ради з наукових і технологічних досліджень Турецької Республіки (TUBITAK) в рамках міжнародного українсько-турецького науково-технічного проекту №М/47-2008 "Розробка методів проектування та оптимізації систем виявлення порушників безпеки".

4. ВИСНОВКИ

Розроблено мережевий контролер та мережевий адаптер, які дозволили, в сукупності зі стандартними сповіщувачами безпеки, створити дистрибутивну сенсорну мережу з топологією "спільна шина". Перевагою запропонованого рішення є зменшення витрат кабелю та вартості обладнання систем безпеки. При цьому розроблені мережевий контролер і мережевий адаптер мають нескладну конструкцію з використанням дешевих компонентів.

5. СПИСОК ЛІТЕРАТУРИ

[1] Р.Г. Магауенов. *Системы охранной сигнализации: основы теории и принципы построения. Учебное пособие.* – М.: Горячая линия, 2004. – 367 с.

[2] *Perimeter Security Sensor Technologies Handbook*. Electronic Security Systems Engineering Division. – North Charleson. – South Carolina, 1997, p. 107 (URL http://www.nlectc.org/perimetr/Hb-Word.doc).

[3] I. Turchenko, V. Turchenko, V. Kochan, P. Bykovyy, A. Sachenko, G.Markowsky. Database design for CAD system optimising distributed sensor networks for perimeter security. *Proceedings of the 8th IASTED International Conference Software Engineering and Applications*, November 9-11, 2004, MIT, Cambridge, MA, USA, pp. 59-64.

[4] P. Bykovyy, V. Kochan, A. Sachenko, G. Markowsky. A CAD System that optimizes distributed sensor networks for perimeter security. *Proceedings of the Second IEEE International Conference on Technologies for Homeland Security and Safety*, Istanbul, Turkey, October 9-13, 2006. pp. 271-276.

[5] ДСТУ IEC 60839-2-2-2001. Системи тривожної сигналізації. Частина 2. Вимоги до систем охоронної сигналізації. Розділ 2. Вимоги до сповіщувачів. Загальні принципи. *Держстандарт України*, Київ. 2002.

[6] С. Звежинский, В. Иванов, В. Рудниченко. Классификации, особенности и информационно – измерительные модели средств обнаружения. *Специальная техника*, № 6, 2007.

[7] Crow Electronic Engineering, Inc, http://www.crowelec.com/, 2160 North Central Road, Fort Lee, NJ 07024 USA.

[8] П. Биковий. Апаратні засоби мережі сенсорів систем безпеки. *Збірник тез міжнародної науково-технічної конференції "Комп'ютерні системи та мережні технології"*, Національний авіаційний університет, м. Київ, Україна, 17-19 березня, 2008, с.44-48.

[9] ftp://ftp.elin.ru/pdf/1-Wire/standard.pdf

[10] Патент 25609А України, МКІ G06F 15/00. Двопровідна локальна обчислювальна мережа, повторювач сигналу та інвертор для використання в ній / *В.В.Кочан,*

В.О.Тимчишин (Україна); Заявл. 30.10.97 № 97105295; Видано 30.10.98.

[11] *V. Kochan, V. Tymchyshyn.* Construction of distributed information measurement systems on the basis of modified RS-232C interface. *Proceedings of the 10th IMEKO TC-4 Symposium on Development in Digital Measuring Instrumentation*. Naples, Italy, 1998, pp. 723-726.

[12] www.atmel.com, 2325 Orchard Parkway, San Jose, CA 95131

[13] А.Б.Семенов. *Проектирование и расчет структурированных кабельных систем и их компонентов*. М.: ДМК Пресс; М.: Компания АйТи, 416с.

***Павло Биковий**, аспірант, кафедра інформаційно-обчислювальних систем та управління, Науково-дослідний інститут інтелектуальних комп'ютерних систем, Тернопільський національний економічний університет. Інтереси: системи безпеки, системи автоматизованого проектування.*

computing@computingonline.net
www.computingonline.net

ISSN 1727-6209
International Journal of Computing

DISTRIBUTED SENSOR NETWORK FOR SECURITY SYSTEMS

Pavlo Bykovyy

Research Institute of Intelligent Computer Systems
Ternopil National Economic University
pb@tneu.edu.ua

Abstract: *the low-cost network controller for security systems detectors was designed. The controller's specifics lies in two-wired network interface with the "common bus" topology support. This design reduces the amount of data communication channels from detectors.*

Keywords: *Security systems, network controller, two-wired network.*

1. INTRODUCTION

Security systems and systems that prevent non-authorized access have become tremendously popular [1]. As a result, a great diversity of such systems has been created with many different functional characteristics [2]. This has led to the construction of the complex security systems that defend against many types of threats, and which combine different subsystems into a coherent whole.

Traditionally security systems have a "star" topology, so there is a substantial expense in providing the cables [1, 2]. It is possible to reduce the number of cables and whole price of the system using a network controller which interfaces detectors of security systems. It will allow to create security systems that have a "common bus" topology where informational signals and power supply are provided by two-wired network. Thus, the total length of cables will include the distance between two most remote detectors and sum of branches to each detector.

This paper describes the distributed sensor network (DSN) architecture design based on the security system detectors described below.

2. ARCHITECTURE OF DESIGNED DISTRIBUTED SENSOR NETWORK

Architecture of the designed distributed sensor network consists of a proposed network controller and a network adapter [3]. The network controller should provide following functions:

- output signals processing of different type and operation principle security detectors;
- to support the specialized wired network of security detectors with a "common bus" topology and supply power through the communication lines.

Network adapter should provide data communication and power supply of network controllers and detectors.

The general structure of standard component of the designed DSN based on proposed universal network controller *UNC*, that interfaces the detectors (sensor units, *SU*) of security systems (fig.1), consists of the sensing elements $S_1 \ldots S_n$ of detector *SU* that interfaces the analog processing circuit and processor *(APCP)* which transmit results (detection/non detection) to output circuit sensor unit (*OPSU*). The microcontroller *MC* interrogates output circuit sensor unit outputs and interacts with the network by hardware tool of serial interface *IF HW*. Power supply *PS* provides power to all of the elements.

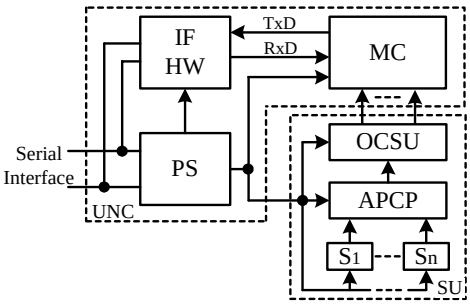


Fig. 1 – Standard component architecture of the developed DSN

It is proposed to use in the developed controller the modified serial RS-232C interface described in [4, 5]. This interface can be supplied with power in the same manner as the 1-Wire interface provided by Dallas Semiconductors [6].

During its work the network server takes turns communicating with the controllers of the security system, sending appropriate requests using its serial interface. Server inquiry goes to the microcontroller *MC* via circuit *IF HW*. *MC* interrogates the outputs of the output circuit sensor unit for detection of sensing elements $S_1 \ldots S_n$, and then forms a response out of the collected data and sends it back to the network server. At the same time *PS* forms from server requests power supply for the controller and detector.

3. DSN IMPLEMENTATION USING SECURITY DETECTORS

Standard component of developed DSN (fig.1) which include security detector SRPG-2N and network controller (fig. 2) can work in two-wired networks with the "common bus" topology that provides power supply for detectors using the network. Total number of detectors is less that 64 where each can have more that one sensing element.

Fig. 2 – Component of DSN based on SRPG-2N detector with connected network controller

Block diagram of DSN (fig. 3) consists of personal computer (network server) that executes the central panel functions and has one or more COM-ports to which the network adapter is connected. The network adapter is connected using two-wired network with each component of DSN.

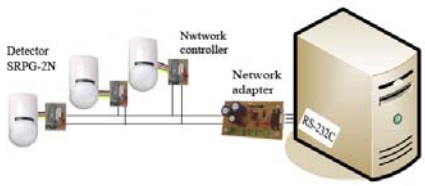


Fig.3. Block diagram of designed DSN.

ACKNOWLEDGMENTS

This work is supported by the Ministry of Education and Science of Ukraine and the Scientific and Technical Research Council of Turkey (TUBITAK) within joint Ukrainian-Turkey project M/47-2008 entitled "Design and Optimization Methods of Physical Intrusion Detection for Security Systems".

4. CONCLUSION

A universal network controller and a network adapter was developed. Together with traditional security detectors it allowed to implement the distributive sensor network with "common bus" topology. The advantage of the proposed solution is reduction of the cabling amount and security system equipment costs. Thus, developed controller and network adapter has a simple design and designed using low-cost components.

5. REFERENCES

[1] R. Magauyenov. *Alarm Security Systems: theory basis and principles of construction.* Teaching aid, Hot line, 2004, 367 p. (in Russian).

[2] *Perimeter Security Sensor Technologies Handbook.* Electronic Security Systems Engineering Division. – North Charleson. – South Carolina, 1997, p. 107 (URL http://www.nlectc.org/perimetr/Hb-Word.doc).

[3] P. Bykovyy. Hardware tools of security sensor networks, *Thesis of International Conference "Computer Systems and Network Technologies"*, National Aviation University, Kiev, Ukraine, 17-19 March, 2008, pp.44-48 (in Ukrainian).

[4] Patent 25609A Ukraine, IPC G06F 15/00. Two-wired local area network, transponder and inverter / V.V. Kochan, V.O. Tymchyshyn (Ukriane); Filled 30.10.97, # 97105295; Issued 30.10.98.

[5] V. Kochan, V. Tymchyshyn. Construction of distributed information measurement systems on the basis of modified RS-232C interface. *Proceedings of the 10th IMEKO TC-4 Symposium on Development in Digital Measuring Instrumentation.* Naples, Italy, 1998, pp. 723-726.

[6] ftp://ftp.elin.ru/pdf/1-Wire/standard.pdf