НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

UDC 004.93

THE FRACTAL ANALYSIS OF SAMPLE AND DECISION TREE MODEL

Subbotin S. A. – Dr. Sc., Professor, Head of the Department of Software Tools, National University "Zaporizhzhia Polytechnic", Zaporizhzhia, Ukraine.

Gofman Ye. A. – PhD, Senior Researcher of the Research Unit, National University "Zaporizhzhia Polytechnic", Zaporizhzhia, Ukraine.

ABSTRACT

Context. The problem of decision tree model synthesis using the fractal analysis is considered in the paper. The object of study is a decision trees. The subject of study is a methods of decision tree model synthesis and analysis.

Objective. The objective of the paper is a creation of methods and fractal indicators allowing jointly solving the problem of decision tree model synthesis and the task of reducing the dimension of training data from a unified approach based on the principles of fractal analysis.

Method. The fractal dimension for a decision tree based model is defined as for whole training sample as for specific classes. The method of the fractal dimension of a model based on a decision tree estimation taking into account model error is proposed. It allows to built model with an acceptable error value, but with optimized level of fractal dimensionality. This makes possibility to reduce decision tree model complexity and to make it mo interpretable. The set of indicators characterizing complexity of decision tree model is proposed. The set of indicators characterizing complexity of decision tree model is proposed. It contains complexity of node checking, complexity of node achieving, an average model complexity and worst tree model complexity of computations. On the basis of proposed set of indicators a complex criterion for model building is proposed. The indicators of the fractal dimension of the decision tree model error can be used to find and remove the non-informative features in the model.

Results. The developed indicators and methods are implemented in software and studied at practical problem solving. As results of experimental study of proposed indicators the graphs of their dependences were obtained. They include graphs of dependencies of number of hyperblocks covering the sample in the features space from size of block side: for whole sample, for each class, for different set error values and obtained error values, for varied values of resulted number of features and instances, also as graphs of dependencies between average and worst tree complexities, decision tree fractal dimensionality and tree average complexity, joint criterion and indicator of feature set reduction, and between joint criterion and tree fractal dimensionality/

Conclusions. The conducted experiments confirmed the operability of the proposed mathematical support and allow recommending it for use in practice for solving the problems of model building by the precedents.

KEYWORDS: decision tree, sample, fractal dimension, indicator, tree complexity.

NOMENCLATURE

ε is a maximum acceptable error value;

ω is a set of model parameters;

c_i^a is a complexity of achieving of *i*-th leaf node;

c_i is a complexity of checking for *i*-th node it's can be obtained as number of *i*-th node's successors;

c_{tree}^w is a worse complexity of computations for the tree model;

c_{tree}^a is an average complexity of computations for the tree model;

D is a fractal dimension;

$<Dc>$ is a correlation dimension;

D_{tree} is a data fractal dimension relatively the accuracy (error) of the synthesized model;

$D^{(k)}$ is a fractal dimension of *k*-th class;

D is a fractal dimension of the sample;

E is a model error;

f is a model quality criterion;

$F()$ is a model structure;

F_{tree} is a joint multiplicative criterion for decision tree model;

I_N is a coefficient of features reduction;

j is a number of feature;

K is a number of classes;

DOI 10.15588/1607-3274-2020-1-11

K is a number of classes;

L a number of intervals on which the ranges of feature values will be separated;

l is a hypercube side length;

l is a length of the interval;

L is a number of intervals;

N is an number of input features;

N' is a feature subset size;

$n(l)$ is a number of hyperblocks of side with the *l* size covering the sample;

$n_{i,q}$ is a number of instances belonging to a rectangular hyperblock formed by feature intervals;

$n_{i,q,k}$ is a number of *k*-th classes exemplars, which are given in each rectangular hyperblock formed by features intervals;

$n(l)$ is a number of hyperblocks with the side of *l* size covering the sample;

$n_k(l)$ is a number of hyperblocks with the side of *l* size covering the sample for *k*-th class;

$n_{(k)}$ is a number of hyperblocks with the side of *l*-size covering the *k*-th class of the sample;

opt is a symbol of optimum;

Q is a number of clusters;

r is a heuristically defined cut-off radius;

r_k is a Euclidean distance between pair of points;

S' is a subsample size;

S is a number of precedents;

tree is a tree recognizing model;

U is a total number of tree nodes;

u_i is a type of the *i*-th node of the tree;

X is a data sample;

x_j is a *j*-th input feature;

x^s is a *s*-th instance of a sample;

x^s_j is a value of *j*-th input for *s*-th instance;

$x_j^{\max}$ is a maximal value of x_j;

$x_j^{\max}$ is a minimal value of x_j;

Y^i is a class of majority of instances hit to the *i*-th node, which is a leaf;

y is an output feature vector;

y^s is a value of output feature values for *s*-th instance.

INTRODUCTION

Decision trees are a popular tool for solving problems of building models on precedents in diagnostics, pattern recognition, and forecasting in various practical areas [1–4]. One of the most significant advantages of models based on decision trees is their interpretability (convenience for human perception and analysis).

The **object of study** is a decision trees.

It is now known a large number of methods to synthesize a model based on decision trees [5–11]. However, as a rule, the known methods in their goal functions (the criteria for the training quality) do not take into account the characteristics of the training sample. This in practice can lead to the construction of non-optimal models.

On the other hand, the model synthesis for big data sets unusually requires the preliminary reduction of the data dimensionality size, which is explained by the high iterativity of the known training methods, as well as the need to obtain a model that provides a good generalization of the data. At the same time, the traditionally used methods of informative feature selection [12–15] and of sample formation [16–21] have such common disadvantage as they are not directly related to each other and come from different points of view on the informativeness of features or instances.

The **subject of study** is methods of decision tree model synthesis and analysis.

One of the promising areas of data analysis is a fractal analysis [22–31]. There are various approaches to the definition of fractal parameters for data [25, 27]. However, they are also not interconnected with each other and with the decision tree model training process.

The **objective** of the paper is a creation of methods and fractal indicators allowing jointly solving the problem of decision tree model synthesis and the task of reducing the dimension of training data from a unified approach based on the principles of fractal analysis.

1 PROBLEM STATEMENT

Let we have an original data sample $X = \langle x, y \rangle$ a set of *S* precedents (instances. exemplars, observations) characterizing dependence $y(x)$, where $x=\{x^s\}$, $y=\{y^s\}$, $s=1, 2, \ldots S$, characterized by the set of *N* input features $\{x_j\}$, $j = 1, 2, \ldots, N$, and output feature *y*. Each *s*-th precedent can be noted as $\langle x^s, y^s \rangle$, $x^s=\{x^s_j\}$, $y^s \in \{1, 2, \ldots, K\}$, $K>1$.

Then the problem of model synthesis of the dependence $y(x)$ will be considered in a search of such structure $F()$ and adjusting such values of parameters ω of a model $\langle F(), \omega \rangle$, which will satisfy the model quality criterion $f(F(), \omega, \langle x, y \rangle) \rightarrow opt$. Usually, the model quality criterion is defined as a model error [2]:

$$E = \frac{1}{2}\sum_{s=1}^{S}(y^s - F(\omega, x^s))^2 \rightarrow \min.$$

2 REVIEW OF THE LITERATURE

The key concept in a fractal analysis is a fractal dimension, which is defined as coefficient describing the fractal structure or the set on the basis of a quantitative assessment of its complexity as the coefficient of variation in details and with a scale conversion.

The Hausdorff-Besicovich dimension according to [25, 28] is defined as

$$D \approx \frac{\log(n(l))}{\log\left(\frac{1}{l}\right)}.$$

One of the most affordable ways to determine the Hausdorff-Besicovich dimension is box-counting method [29, 30], which consists in repeating fractal object coating

DOI 10.15588/1607-3274-2020-1-11

by hypercubes of equal size and counting minimum number of hypercubes which contain points of the object.

By consistently reducing the hypercubes size l we will get a set of points with coordinates $(\log(n(l)),\ \log(l^{-1}))$, which define a curve, which slope determined by the linear regression, is a fractal dimension:

$$D = \lim_{l \to 0} \frac{\log(n(l))}{\log\left(\frac{1}{l}\right)}.$$

The Takens' method [31] is used to determine the correlation dimension:

$$\langle Dc \rangle = -\left\{ \frac{1}{|R|} \sum_{k=1}^{|R|} r_k \right\}^{-1},$$

where $R = \{r_k |\ r_k < r\}$, $|R|$ is a cardinality of the set R, $r>0$.

The common disadvantage of considered methods [22–31] for determining the fractal dimension is that the cardinality of the set must satisfy the inequality $N<2\log_{10}S$, which shows that the number of data points S required for accurate dimension estimation of the N-dimensional set must be at least $10^{\frac{N}{2}}$. It leads to large N values even for small sets.

The common feature of all above described methods for determining the fractal dimension is that sample dimension and dimension of the model trained on its basis are defined with no connection to each other. It limits their practical application.

In [32] the methods for estimating the fractal dimension allowing characterize properties of the sample. The sample instances are represented as points in the feature space. Then clusters will correspond to the compact areas in a feature space, which will be combined into classes. Different geometric shapes can describe clusters. Fractal analysis of the sample in the feature space can be performed by setting the elemental form for clustering and varying the size of the cluster for partitioning the sample into fragments. For the sample fractal dimension analysis the method [32] contains following stages.

Initialization stage. Set a learning sample $<x,y>$ and L the number of intervals on which the ranges of feature values will be separated.

Sample normalization stage. If feature values are non-normalized, they should be normalized by mapping to the interval [0, 1]: $x_j^s = \left(x_j^s - x_j^{\min}\right) / \left(x_j^{\max} - x_j^{\min}\right)$.

Clustering stage. Divide the range of each feature values on L intervals of length l: $l = 1/L$. Form clusters as rectangular blocks at the different features interval intersection.

Data analysis stage. Determine the number of instances belonging to a rectangular hyperblock formed by feature intervals $n_{i,q}$. Determine the number of k-th classes exemplars, which is given in each rectangular hyperblock formed by features intervals $n_{i,q,k}$.

Determine the number of hyperblocks with the side of l-size covering the k-th class of the sample in the N features space:

$$n_{(k)} = \sum_{i=1}^{N} \sum_{q=1}^{L} \left\{ 1 \mid n_{i,q,k} > 0 \right\}.$$

Determine the number of hyperblocks with the side of l size covering the sample in the N features space:

$$n(l) = \sum_{i=1}^{N} \sum_{q=1}^{L} \left\{ 1 \mid n_{i,q} > 0 \right\} = \sum_{i=1}^{N} \sum_{q=1}^{L} \left\{ 1 \left| \sum_{k=1}^{K} n_{i,q,k} > 0 \right. \right\}.$$

Stage of fractal dimension estimation. Determine at a given l the fractal dimension of k-th class, $k = 1, 2, ..., K$: $D^{(k)} = \log(n_{(k)}) / \log(l^{-1})$.

Determine the fractal dimension of the sample at a given l: $D = \log(n(l)) / \log(l^{-1})$.

This method operates with rectangular blocks of the same size, covering the feature space by them. The single controlled parameter of the method is defined by the number of intervals L, which are divided in ranges of the feature values.

It is obvious that number of clusters $Q \geq K$, $Q=L^N$, and for each feature $L \geq 2$. To provide generalization properties of clusters we impose restriction $Q \leq NS$.

Thus, we obtain $K \leq L^N \leq NS$, $L \geq 2$. Taking logarithms $\log(K) \leq N\log(L) \leq \log(NS)$ we obtain after transformations: $2 \leq L \leq \sqrt[N]{NS}$. Note that minimum step for varying the L values is 1. If the upper limit value $\sqrt[N]{NS}$ is less than 2, it can be replaced with S. This is due to the fact that on the each feature axis will not be more than the S points and feature axis partition on more than S intervals will obviously lead to occurrence of the empty intervals. For large N values, the given number of partitions S on each feature will lead to forming of a huge number of blocks equal S^N, which make a computation very hard, and in some cases practically non-realizable. Therefore, it is reasonable in this case to set the value of the upper limit of L by the round(log(S)), where round is function of rounding to the nearest integer number. Evaluation of indicator D for small values of L requires high cost of computing resources and computer memory resources than for large L values. However, the analysis accuracy for small L values will be lower while the generalization level will be higher than for large L values.

Consider possible ways to implement this method. If we assume that the data structure will be created containing the counters of instances numbers belonging to each of rectangular hyper-block in the feature space, it will require at least $2L^N$ memory cells where 2 bytes will be given to represent L^N integers. In turn, for each hyper-block we need to evaluate belonging of the sample instances, which would require about $2SL^N$ comparisons. This approach, obviously, is practically applicable only for small N. Since to determine the fractal dimension it is not important to know how many instances hit in each

DOI 10.15588/1607-3274-2020-1-11

block, but it is important to know how many blocks contains instances, then to reduce the computational and memory costs are encouraged to use the following approach.

The advantage of the described method and of the sample quality indicator determined on its basis is the fact that they does not depend of the model synthesis method, and of the results of its work and allow to evaluate the properties of the single sample.

The disadvantages of this method are the uncertainty in the choice of the L parameter value, and absence of relation between the method and the quality of the synthesized model.

3 MATERIALS AND METHODS

The decision tree model consists of nodes connected by the links. The node can be a root (having no parents), a leaf (having no successors), or an internal (having parent and successors nodes). Each node of the tree (excluding leafs) contains check on one of the features. As a result of checking the recognized instance on this node, it will be redirected to one of the successor nodes of this node, depending on what interval of checked feature values it falls into.

For a decision tree based model, we define the fractal dimension as the minimum number of rectangular blocks in the feature space needed to cover the training data set. Since the leaf nodes of the model based on the decision tree correspond to rectangular areas in the feature space, and the instances of the training sample belong by the model only to these areas, the number of leaf nodes in the tree is the fractal dimension of the decision tree.

Let u_i is a type of the i-th node of the tree (u_i = 1 if i-th node is a leaf; u_i = 0, otherwise), U is a total number of tree nodes. Then the number of hyperblocks with the side of l size covering the sample in the normalized feature space can be evaluated as:

$$n(l) = \sum_{i=1}^{U} u_i \ .$$

By analogy for k-th class in the sample we can define:

$$n_k(l) = \sum_{i=1}^{U} \left\{ 1 \mid u_i = 1, Y^i = k \right\}, \ k = 1, 2, ..., K,$$

where Y^i is a class of majority of instances hit to the i-th node, which is a leaf, K is a number of classes.

It is obviously, that

$$n(l) = \sum_{k=1}^{K} n_k(l) \ .$$

To estimate the fractal dimension of a model based on a decision tree, we will use an approach similar to neural networks [33–35].

Initialization stage. Set the training sample <x, y>, the model synthesis method, the model training quality criterion as error function E, and the maximum acceptable error value ε.

Sample normalization stage. If feature values are non-normalized, they should be normalized by mapping to the interval [0, 1].

Formation and analysis of data partition stage. Sequentially changing the value of L = 2, ..., S:

– determine the length of the interval l;

– quantize the sample features, partitioning their ranges of values on L intervals;

– determine the number of hyperblocks of side with the l size covering the sample in the space of the N features $n(l)$;

– prune a recognizing model *tree* by a given method, minimizing the error function E to achieve an acceptable level ε;

– estimate the error E of the constructed recognizing model *tree*.

The fractal dimension determining stage. For every l, for which the model error E is acceptable, determine the data fractal dimension relatively the accuracy (error) of the synthesized model *tree*:

$$D_{tree} = \left\{ \left. \frac{\log(n(l))}{\log\left(\frac{1}{l}\right)} \right| E(tree) \le \varepsilon \right\}.$$

This method operates by rectangular blocks of equal size, covering by them the feature space. The single controlled parameter of the method is given threshold value of the model error ε.

Obviously, the smaller the given ε value, the more detailed model should be, i.e., it will need to form a larger number of clusters Q, and hence the greater should be the L value. Accordingly, with a decrease of the given ε, the cost of computing resources and computer memory resources will increase for the sample analysis.

The advantage of the proposed method and sample quality indicator determined on its basis is the fact that they are related with quality indicator of the synthesized model, and automatically sets the optimum value of the L.

The disadvantages of the proposed method are the uncertainty of ε parameter values choice and its dependence on the training quality and model functioning principles on which it is defined. It should also be noted that the error function used in the method is only one of the synthesized model characteristics, but it does not take into account the model dimension and generalizing properties.

Therefore, the fractal dimension of the trained model is proposed to be determined on the basis of the below method taking into account the model dimension.

In addition to the tree fractal dimensionality we can take into account complexity of calculations.

For i-th node it's complexity of checking c_i can be obtained as number of i-th node's successors.

DOI 10.15588/1607-3274-2020-1-11

For *i*-th leaf node the complexity of achieving c_i^a can be evaluated as a sum of complexities of checking of all nodes in the path from the tree root to the *i*-th leaf node.

For the tree model the worse complexity of computations can be estimated as the maximal complexity of the leaf nodes:

$$c_{tree}^w = \max_{i=1,2,...,U} \{c_i^a \mid u_i = 1\}.$$

For the tree model the average complexity of computations can be estimated as the average complexity of the leaf nodes:

$$c_{tree}^a = \frac{1}{n(l)} \sum_{i=1}^{U} \{c_i^a \mid u_i = 1\}.$$

Generally, when the model error is acceptable, we need to minimize the average computational complexity of leaf nodes reaching, as well as minimize the worst complexity of leaf nodes reaching.

For the decision tree model synthesis the proposed set of fractal indicators allows to define a system of criterions:

$$\begin{cases} D_{tree} \to \min \\ c_{tree}^a \to \min \\ c_{tree}^w \to \min \end{cases}.$$

Since in the general case the number of features in the original set and the number of features used by the model based on the decision tree may differ, it is advisable to consider it as a characteristic of the model quality. Then it is possible to determine on its basis the coefficient of features reduction as:

$$I_N = \frac{N}{N'}, 1 \le N' \le N.$$

Obviously, the greater the coefficient of feature reduction, the simpler the model, provided that acceptable accuracy is achieved. In the best case $I_N = N$, in the worst case $I_N = 1$.

It is possible also to define one joint multiplicative criterion for decision tree model synthesis based on the fractal analysis as:

$$F_{tree} = \frac{D_{tree} c_{tree}^a c_{tree}^w}{I_N} \to \min.$$

Using proposed fractal indicators as individual, and as combined it is possible to solve different tasks in the process of decision tree model synthesis from unified point of view.

4 EXPERIMENTS

To study the complex of proposed sample and model fractal indicators they were implemented in software. The developed software was used in computational experiments to study the applicability of proposed indicators for solving the problems of automatic classification.

Several datasets for different tasks [4, 36–38] characterized in Table 1 were used for experimental study.

Table 1 – Characteristics of the tasks for methods experimental study

Task	Source	*N*	*S*	*K*
Fisher Iris	[36]	4	150	3
Agricultural plant classification on remote sensing data	[37]	55	248	2
Diagnosis of Arrhythmia	[38]	279	452	2
Air-engine blade diagnosis	[4]	10240	32	2

For this datasets several series of experiments were conducted.

The first series of experiments were devoted to study the methods of data dimensionality reduction using fractal indicators for model synthesis. Here it is needed to evaluate fractal dimensionalities of original datasets and their classes. Then is possible to study dependencies of $n(l)$ from l^{-1} for the entire sample and the classes, for different given ε values and obtained values of error E, as well as dimensions of the formed data subsample: subsample size S' and feature subset size N'.

The second series of experiments were concerned to study the methods of decision tree model synthesis using fractal indicators. For each task we need to built a tree model and study dependencies between sample properties and proposed indicators.

5 RESULTS

For each data set as a result of the experiments, the fractal dimensions of the data and the decision tree models constructed on their basis are calculated.

For example, the computed fractal dimension of the sample for the Fisher Iris data set [36] is $D = 0.59034$, and fractal dimension assessments of the classes: $D^{(1)} = 0.68223$, $D^{(2)} = 0.6212$, $D^{(3)} = 0.53407$.

Graphs of dependencies from l^{-1} in a logarithmic system of coordinates for the entire sample and the classes are shown in Fig. 1 and Fig. 2, respectively. On Fig. 2 markers of different sizes encode the different classes.

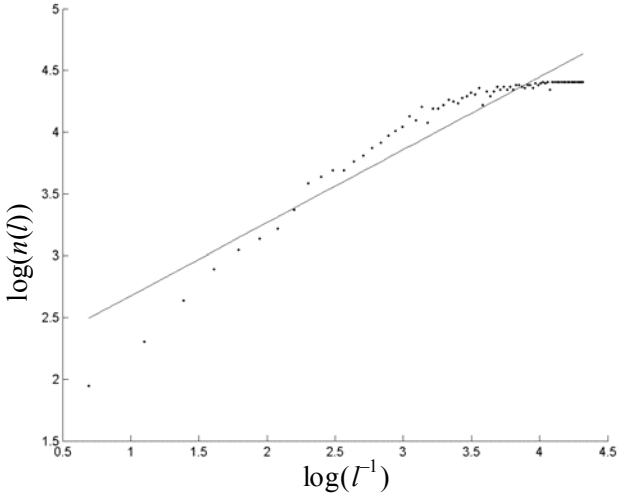


Figure 1 – Graph of dependency of $n(l)$ of a sample from from l^{-1} in a logarithmic coordinate system

DOI 10.15588/1607-3274-2020-1-11

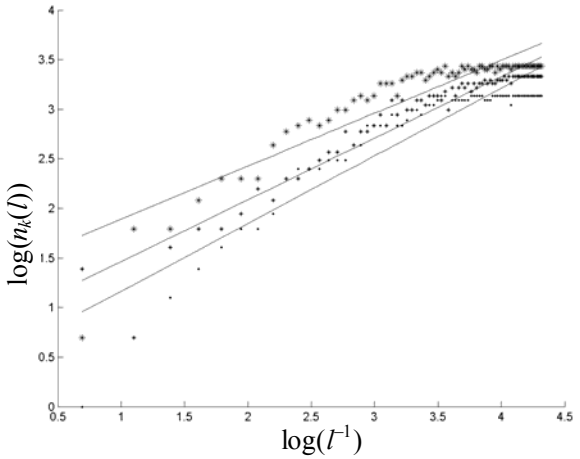


Figure 2 – Graphs of dependencies of $n_k(l)$ of classes from l^{-1} in a logarithmic coordinate system

Fig. 3 shows the schematic graph of generalized dependencies of $n(l)$ of sample from l^{-1} in a logarithmic system of coordinates for different set values of ε and obtained values E.

Fig. 4 presents the schematic graph of generalized dependency of $n(l)$ from l^{-1} in a logarithmic coordinate system for varied values of N' and S'.

Fig. 5 shows the schematic graph of generalized dependencies between c^a_{tree} and c^w_{tree}.

Fig. 6 presents the schematic graph of generalized dependencies between D_{tree} and c^a_{tree}.

Fig. 7 shows the schematic graph of generalized dependency between F_{tree} and I_N.

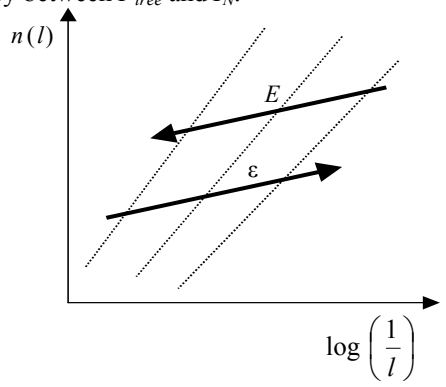


Figure 3 – Schematic graph of generalized dependency of $n(l)$ from l^{-1} in a logarithmic coordinate system

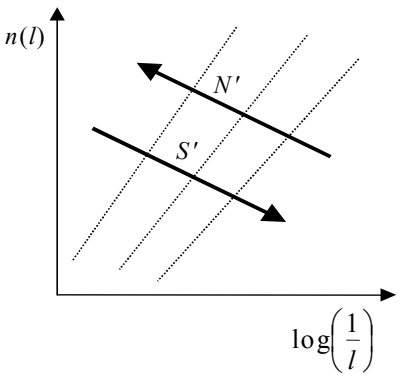


Figure 4 – Schematic graph of generalized dependency of $n(l)$ from l^{-1} in a logarithmic coordinate system

DOI 10.15588/1607-3274-2020-1-11

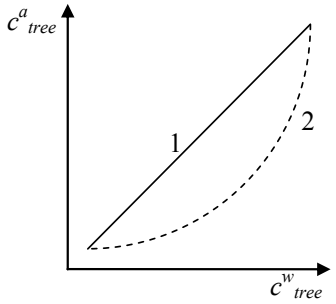


Figure 5 – Schematic graph of generalized dependencies between c^a_{tree} and c^w_{tree}: 1 – $c^a_{tree} = c^w_{tree}$, 2 – $c^a_{tree} < c^w_{tree}$

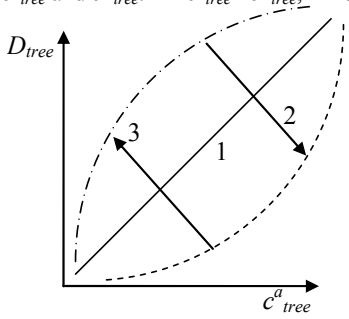


Figure 6 – Schematic graph of generalized dependencies between D_{tree} and c^a_{tree}:
1 – basic relation; 2 – for decreasing of l or $n(l)$ or U or c^w_{tree};
3 – for increasing of l or $n(l)$ or U or c^w_{tree}

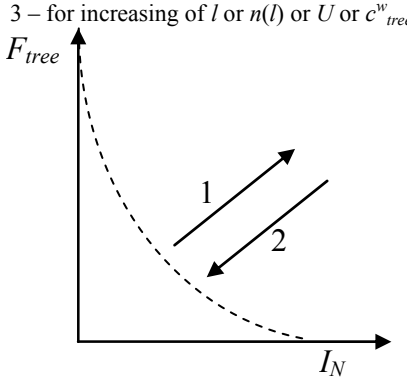


Figure 7 – Schematic graph of generalized dependency between F_{tree} and I_N :
1 – for increasing of model complexity of computations (c^a_{tree} and/or c^w_{tree}), 2 – for decreasing of model complexity of computations (c^a_{tree} and/or c^w_{tree})

Fig. 8 presents the schematic graph of generalized dependencies between F_{tree} and D_{tree}.

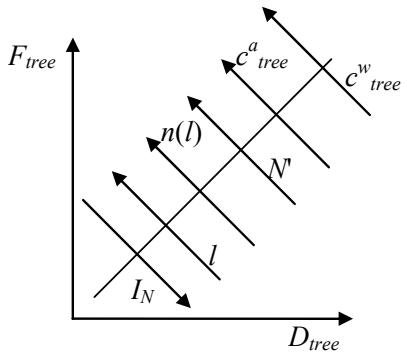


Figure 8 – Schematic graph of generalized dependency between F_{tree} and D_{tree}

6 DISCUSSION

As it can be seen from Fig. 1 and Fig. 2 the proposed indicators of the fractal dimension allow show the differences between classes. These indicators can be used in methods of sample selection, defining quality criteria of formed subsamples on the base of the proposed indicators of the fractal dimension.

If formed subsample or its classes on indicators of the fractal dimension are significantly differ from similar parameters of the original sample, it is possible that, the sample does not have the representativeness relative to the original sample. Also the proposed indicators at the several subsamples-candidates comparing could be used as their quality measures: among subsamples-candidates should be preferred that which have indicators of the fractal dimension with closest values to the original sample indicators values.

As it can be seen from Fig. 3, a change of the specified components of formed subsample dimension (number of features N' and the number of instances S'), also as E and ε affect the position of the straight line connecting points of dependence $n(l)$. The greater the ε value the greater the $n(l)$ and the less the E value the greater the $n(l)$.

From the Fig. 4 we can see that the less S' value and the bigger the N' value the bigger the n(l) value,

As it can bee seen from thr Fig. 5 the bigger the c^{w}_{tree} the bigger the c^{a}_{tree}. The smaller c^{a}_{tree} comparing with c^{w}_{tree} the slower c^{a}_{tree} grow.

Fig. 6 indicates that the bigger c^{a}_{tree} value the bigger the D_{tree} value. If we have decreasing of l or $n(l)$ or U or c^{w}_{tree}; then D_{tree} value will grow slowly comparing with increasing of l or $n(l)$ or U or c^{w}_{tree}

From the Fig. 7 we can bring that the bigger the I_N value the less the F_{tree} value. If model complexity of computations (c^{a}_{tree} and/or c^{w}_{tree}) increased then the bigger F_{tree} and vice versa.

As it can be seen from the Fig. 8 the Ftree indicator will receive the greater value the greater the D_{tree}, l, $n(l)$, N', c^{a}_{tree}, c^{wt}_{ree} and the less the I_N indicator value.

CONCLUSIONS

The urgent problem of decision tree model synthesis using the fractal analysis is considered in the paper.

The scientific novelty of obtained results is that the fractal dimension for a decision tree based model is defined as for whole training sample as for specific classes. The method of the fractal dimension of a model based on a decision tree estimation taking into account model error is proposed. It allows to build model with an acceptable error value, but with optimized level of fractal dimensionality. This makes possibility to reduce decision tree model complexity and to make it mo interpretable.

The set of indicators characterizing complexity of decision tree model is proposed. It contains complexity of node checking, complexity of node achieving, an average model complexity and worst tree model complexity of computations. On the basis of proposed set of indicators a complex criterion for model building is proposed. The indicators of the fractal dimension of the decision tree model error can be used to find and remove the non-informative features in the model.

The practical significance of obtained results is that the developed indicators and methods are implemented in software and studied at practical problem solving. The conducted experiments confirmed the operability of the proposed software and allow recommending it for use in practice for solving the problems of model building by the precedents.

The **prospects for further study** may include the optimization of software implementation of proposed methods and indicators, also as experimental study of proposed indicators on the larger complex of practical problems having different nature and dimension.

ACKNOWLEDGEMENTS

The work was conducted in the framework of the state budget scientific project “Development and research of intelligent methods and software for diagnosing and non-destructive quality control of military and civilian equipment” (State register No. 0119U100360) of National University "Zaporizhzhia Polytechnic" under partial support of international project "Innovative Multidisciplinary Curriculum in Artificial Implants for Bio-Engineering BSc/MSc Degrees” (BIOART, Ref. no. 586114-EPP-1-2017-1-ES-EPPKA2-CBHE-JP)" co-funded by the Erasmus+ Programme of the European Union and "Virtual Master Cooperation on Data Science (ViMaCs) funded by the DAAD.

REFERENCES

1. Geurts P., Irrthum A., Wehenkel L. Supervised learning with decision tree-based methods in computational and systems biology, *Molecular Biosystems*, 2009, Vol. 5, No. 12, pp. 1593–1605.
2. Breiman L., Friedman J. H., Olshen R. A., Stone C. J. Classification and regression trees. Boca Raton, Chapman and Hall/CRC, 1984, 368 p.
3. Heath D., Kasif S., Salzberg S. Induction of oblique decision trees [Electronic resource]. Baltimor, Johns Hopkins University, 1993, 6 p. Access mode: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.9208&rep=rep1&type=pdf
4. Rabcan J., Levashenko V., Zaitseva E., Kvassay M., Subbotin S. Non-destructive diagnostic of aircraft engine blades by Fuzzy Decision Tree, *Engineering Structures.* – Vol. 197, 109396.
5. Quinlan J. R. Induction of decision trees, *Machine learning*, 1986, Vol. 1, No. 1, pp. 81– 106.
6. Breiman L. Bagging predictors , *Machine Learning*, 1996, Vol. 24, No. 2, pp. 123–140.
7. Utgoff P. E. Incremental induction of decision trees, *Machine learning*, 1989, Vol. 4, No. 2, pp. 161–186. DOI:10.1023/A:1022699900025
8. Hyafil L., Rivest R. L. Constructing optimal binary decision trees is np-complete, *Information Processing Letters.* – 1976, Vol. 5, No. 1, pp. 15–17.
9. Subbotin S. A. Postroyeniye derev’yev resheniy dlya sluchaya maloinformativnykh priznakov, *Radío Electronics, Computer Science, Control*, 2019, No. 1, pp. 122–131.

DOI 10.15588/1607-3274-2020-1-11

10. Amit Y., Geman D., Wilder K. Joint induction of shape features and tree classifiers, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997, Vol. 19, No. 11. pp. 1300–1305.
11. Subbotin S. A. Metody sinteza modeley kolichest-vennykh zavisimostey v bazise derev'yev regressii, realizuyushchikh klaster-regressionnuyu approksimatsiyu po pretsedentam, *Radío Electronics, Computer Science, Control,* 2019, No. 3, pp. 76-85.
12. Jensen R., Shen Q. Computational intelligence and feature selection: rough and fuzzy approaches. Hoboken, John Wiley & Sons, 2008, 300 p.
13. Subbotin S. The instance and feature selection for neural network based diagnosis of chronic obstructive bronchitis, *Applications of Computational Intelligence in Biomedical Technology*. Cham, Springer, 2016, pp. 215–228. DOI: 10.1007/978-3-319-19147-8_13
14. Miyakawa M. Criteria for selecting a variable in the construction of efficient decision trees, *IEEE Transactions on Computers*, 1989, Vol. 38, № 1, pp. 130–141.
15. Tolosi L., Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions, *Bioinformatics,* 2011, Vol. 27, No. 14, pp. 1986–1994. DOI:10.1093/bioinformatics/btr300
16. Chaudhuri A., Stenger H.. Survey sampling theory and methods. New York, Chapman & Hall, 2005, 416 p. DOI: 10.1201/9781420028638
17. Subbotin S.A. Methods of sampling based on exhaustive and evolutionary search, *Automatic Control and Computer Sciences,* 2013, Vol. 47, No. 3, pp. 113–121. DOI: 10.3103/s0146411613030073
18. Subbotin S.A. The sample properties evaluation for pattern recognition and intelligent diagnosis, *Digital Technologies : 10th International Conference, Zilina, 9–11 July 2014 : proceedings*. Los Alamitos, IEEE, 2014, pp. 332–343. DOI: 10.1109/dt.2014.6868734
19. Lavrakas P.J. Encyclopedia of survey research methods. – Thousand Oaks: Sage Publications, 2008, Vol. 1–2, 968 p. DOI: 10.4135/9781412963947.n159
20. Łukasik S., Kulczycki P. An algorithm for sample and data dimensionality reduction using fast simulated annealing, *Advanced Data Mining and Applications, Lecture Notes in Computer Science*. Berlin: Springer, 2011, Vol. 7120, pp. 152–161. DOI: 10.1007/978-3-642-25853-4_12
21. Subbotin S. A. The training set quality measures for neural network learning, *Optical Memory and Neural Networks (Information Optics),* 2010, Vol. 19, No. 2, pp. 126–139. DOI: 10.3103/s1060992x10020037
22. Cheng Q. Multifractal Modeling and Lacunarity Analysis, *Mathematical Geology*, 1997, Vol. 29, No. 7, pp. 919–932. DOI:10.1023/A:1022355723781
23. Eftekhari A. Fractal Dimension of Electrochemical Reactions, *Journal of the Electrochemical Society,* 2004, Vol. 151, No. 9, pp. E291–E296. DOI:10.1149/1.1773583.
24. Dubuc B., Quiniou J., Roques-Carmes C., Tricot C., Zucker S. Evaluating the fractal dimension of profiles, *Physical Review,* 1989. – Vol. 39, No. 3. – P. 1500–1512. DOI:10.1103/PhysRevA.39.1500
25. Camastra F. Data Dimensionality Estimation Methods: A survey, *Pattern Recognition*, 2003, Vol. 36, No. 12, pp. 2945–2954. DOI: 10.1016/S0031-3203(03)00176-6
26. de Sousa P. M., Traina C., Traina A. J. M., Wu L., Faloutsos C. A fast and effective method to find correlations among attributes in databases, *Data Mining and Knowledge Discovery*, 2007, Vol. 14, Issue 3, pp. 367–407. DOI: 10.1007/s10618-006-0056-4
27. Roberts A., Cronin A. Unbiased estimation of multi-fractal dimensions of finite data sets, *Physica A: Statistical Mechanics and its Applications*, 1996, Vol. 233, No. 3–4, pp. 867–878. DOI:10.1016/s0378-4371(96)00165-3
28. Kumaraswamy K. Fractal Dimension for Data Mining [Electronic resource]. Access mode: https://www.ml.cmu.edu/research/dap-papers/skkumar_kdd_project.pdf.
29. Li J. Du Q., Sun C. An improved box-counting method for image fractal dimension estimation, *Pattern Recognition*, 2009, Vol. 42, No. 11, pp. 2460–2469. DOI:10.1016/j.patcog.2009.03.001.
30. Popescu D. P., Flueraru C., Mao Y., Chang S., Sowa M.G.,Signal attenuation and box-counting fractal analysis of optical coherence tomography images of arterial tissue, *Biomedical Optics Express,* 2010, Vol. 1, No. 1, pp. 268–277. DOI:10.1364/boe.1.000268
31. Takens F. On the numerical determination of the dimension of an attractor, *Dynamical Systems and Bifurcations Workshop Groningen, 16–20 April 1984 : proceedings*. Berlin, Springer, 1985, pp. 99–106. DOI: 10.1007/bfb0075637
32. Subbotin S. A. Metriki kachestva vyborok dannykh i modeley zavisimostey, osnovannyye na fraktal'noy razmernosti, *Radío Electronics, Computer Science, Control*, 2017, No. 2, pp. 70–81.
33. Zong-Chang Y. Establishing structure for artificial neural networks based-on fractal, *Journal of Theoretical and Applied Information Technology*, 2013, Vol. 49, No. 1, pp. 342–347.
34. Crişan D. A., Dobrescu R. Fractal dimension spectrum as an indicator for training neural networks, *Universitatea Politehnica Bucuresti Sci. Bull. Series C,* 2007, Vol. 69, No. 1, pp. 23–32.
35. Subbotin S. A. The neural network model synthesis based on the fractal analysis, *Optical Memory and Neural Networks,* 2017, Vol. 26, No. 4, pp. 257–273.
36. Fisher Iris dataset [Electronic resource]. Access mode: https://archive.ics.uci.edu/ml/datasets/Iris
37. Dubrovin V., Subbotin S., Morshchavka S., Piza D. The plant recognition on remote sensing results by the feed-forward neural networks, *International Journal of Smart Engineering System Design,* 2001, Vol. 3, No. 4, pp. 251–256.
38. Arrhythmia Data Set [Electronic resource]. Access mode: http://archive.ics.uci.edu/ml/datasets/arrhythmia

Received 20.11.2019.
Accepted 16.02.2020.

УДК 004.93

ФРАКТАЛЬНИЙ АНАЛІЗ ВИБІРОК І МОДЕЛЕЙ НА ОСНОВІ ДЕРЕВ РІШЕНЬ

Субботін С. О. – д-р техн. наук, професор, завідувач кафедри програмних засобів Національного університету «Запорізька політехніка», Запоріжжя, Україна.

Гофман Є. О. – канд. техн. наук, старший науковий співробітник науково-дослідної частини Національного університету «Запорізька політехніка», Запоріжжя, Україна.

АНОТАЦІЯ

Актуальність. У статті розглядається проблема синтезу моделі на основі дерева рішень з використанням фрактального аналізу. Об'єктом дослідження є дерева рішень. Предметом дослідження є методи синтезу та аналізу моделей на основі дерев рішень.

DOI 10.15588/1607-3274-2020-1-11

Мета роботи – створення методів і фрактальних індикаторів, що дозволяють спільно вирішити задачу синтезу моделі на основі дерева рішень і завдання скорочення розмірності навчальних даних за допомогою єдиного підходу, заснованого на принципах фрактального аналізу.

Метод. Фрактальна розмірність для моделі на основі дерева рішень визначена як для всієї навчальної вибірки, так і для кожного класу. Запропоновано метод визначення фрактальної розмірності моделі, заснований на оцінюванні дерева рішень з урахуванням похибки моделі. Це дозволяє побудувати модель з прийнятним значенням помилки, але з оптимізованим рівнем фрактальної розмірності, що дозволяє зменшити складність моделі дерева рішень і зробити її більш зрозумілою. Запропоновано набір показників, що характеризують складність моделі на основі дерева рішень. Він містить складність перевірки вузлів, складність досягнення вузла, середню і найгіршу складність обчислень моделі дерева. На основі запропонованого набору показників запропоновано комплексний критерій побудови моделі. Індикатори фрактальної розмірності помилки моделі дерева рішень можуть бути використані для пошуку і видалення неінформативних ознак в моделі.

Результати. Розроблені показники і методи реалізовані в програмному забезпеченні і вивчені при вирішенні практичних завдань. В результаті експериментального дослідження запропонованих показників отримані графіки залежностей між ними, включаючи графіки залежностей числа гіперблоків, що охоплюють вибірку в просторі ознак, від розміру боку блоку: для всієї вибірки, для кожного класу, для різних встановлених значень помилок і отриманих значень помилок, для різних значень результуючих чисел ознак і екземплярів, також графіків залежностей між середньою і найгіршою складнощами дерева, фрактальної розмірністю дерева рішень і ср днів складністю дерева, об'єднаним критерієм і індикатором скорочення набору ознак, а також між спільним критерієм і фрактальної розмірністю дерева.

Висновки. Проведені експерименти підтвердили працездатність запропонованого математичного забезпечення та дозволяють рекомендувати його для практичного використання для вирішення завдань побудови моделей по прецедентах.

КЛЮЧОВІ СЛОВА: дерево рішень, вибірка, фрактальна розмірність, індикатор, складність дерева.

УДК 004.93

ФРАКТАЛЬНЫЙ АНАЛИЗ ВЫБОРОК И МОДЕЛЕЙ НА ОСНОВЕ ДЕРЕВЬЕВ РЕШЕНИЙ

Субботин С. А. – д-р техн. наук, профессор, заведующий кафедрой программных средств Национального университета «Запорожская политехника», Запорожье, Украина.

Гофман Е. А. – канд. техн. наук, старший научный сотрудник научно-исследовательской части Национального университета «Запорожская политехника», Запорожье, Украина.

АННОТАЦИЯ

Актуальность. В статье рассматривается проблема синтеза модели на основе дерева решений с использованием фрактального анализа. Объектом исследования являются деревья решений. Предметом исследования являются методы синтеза и анализа моделей на основе деревьев решений.

Цель работы – создание методов и фрактальных индикаторов, позволяющих совместно решить задачу синтеза модели на основе дерева решений и задачу сокращения размерности обучающих данных с помощью единого подхода, основанного на принципах фрактального анализа.

Метод. Фрактальная размерность для модели на основе дерева решений определена как для всей обучающей выборки, так и для каждого класса. Предложен метод определения фрактальной размерности модели, основанный на оценке дерева решений с учетом погрешности модели. Это позволяет построить модель с приемлемым значением ошибки, но с оптимизированным уровнем фрактальной размерности, что позволяет уменьшить сложность модели дерева решений и сделать ее более понятной. Предложен набор показателей, характеризующих сложность модели на основе дерева решений. Он содержит сложность проверки узлов, сложность достижения узла, среднюю и наихудшую сложность вычислений модели дерева. На основе предложенного набора показателей предложен комплексный критерий построения модели. Индикаторы фрактальной размерности ошибки модели дерева решений могут быть использованы для поиска и удаления неинформативных признаков в модели.

Результаты. Разработанные показатели и методы реализованы в программном обеспечении и изучены при решении практических задач. В результате экспериментального исследования предложенных показателей получены графики зависимостей между ними, включающие графики зависимостей числа гиперблоков, охватывающих выборку в пространстве признаков, от размера стороны блока: для всей выборки, для каждого класса, для различных установленных значений ошибок и полученных значений ошибок, для различных значений результирующих чисел признаков и экземпляров, также графиков зависимостей между средней и наихудшей сложностями дерева, фрактальной размерностью дерева решений и средней сложностью дерева, объединенным критерием и индикатором сокращения набора признаков, а также между совместным критерием и фрактальной размерностью дерева.

Выводы. Проведенные эксперименты подтвердили работоспособность предложенного математического обеспечения и позволяют рекомендовать его для практического использования для решения задач построения моделей по прецедентам.

КЛЮЧЕВЫЕ СЛОВА: дерево решений, выборка, фрактальная размерность, индикатор, сложность дерева.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Geurts P. Supervised learning with decision tree-based methods in computational and systems biology / P. Geurts, A. Irrthum, L. Wehenkel // Molecular Biosystems. – 2009. –Vol. 5, № 12. – P. 1593–1605.
2. Classification and regression trees / [L. Breiman J. H. Friedman, R. A. Olshen, C. J. Stone]. – Boca Raton: Chapman and Hall/CRC, 1984. – 368 p.
3. Heath D. Induction of oblique decision trees [Electronic resource] / D. Heath, S. Kasif, S. Salzberg. – Baltimor : Johns Hopkins University, 1993. – 6 p. – Access mode:

DOI 10.15588/1607-3274-2020-1-11

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.9208&rep=rep1&type=pdf
4. Non-destructive diagnostic of aircraft engine blades by Fuzzy Decision Tree / [J. Rabcan, V. Levashenko, E. Zaitseva et al.] // Engineering Structures. – Vol. 197, 109396.
5. Quinlan J. R. Induction of decision trees / J. R. Quinlan // Machine learning. – 1986. – Vol. 1, № 1. – P. 81–106.
6. Breiman L. Bagging predictors / L. Breiman // Machine Learning. – 1996. – Vol. 24, № 2. – P. 123–140.
7. Utgoff P. E. Incremental induction of decision trees / P. E. Utgoff // Machine learning, 1989. – Vol. 4, № 2. – P. 161–186. DOI:10.1023/A:1022699900025
8. Hyafil L. Constructing optimal binary decision trees is np-complete / L. Hyafil, R. L. Rivest // Information Processing Letters. – 1976. – Vol. 5, № 1. – P.15–17.
9. Субботин С. А. Построение деревьев решений для случая малоинформативных признаков / С. А. Субботин // Радіоелектроніка, інформатика, управління. – 2019. – № 1. – С. 122–131.
10. Amit Y. Joint induction of shape features and tree classifiers / Y. Amit, D. Geman, K. Wilder // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1997. – Vol. 19, № 11. – P.1300–1305.
11. Субботин С. А. Методы синтеза моделей количественных зависимостей в базисе деревьев регрессии, реализующих кластер-регрессионную аппроксимацию по прецедентам / С. А. Субботин // Радіоелектроніка, інформатика, управління. – 2019. – № 3. – С. 76–85.
12. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. –Hoboken: John Wiley & Sons, 2008. – 300 p.
13. Subbotin S. The instance and feature selection for neural network based diagnosis of chronic obstructive bronchitis / S. Subbotin // Applications of Computational Intelligence in Biomedical Technology. – Cham : Springer, 2016. – P. 215–228. DOI: 10.1007/978-3-319-19147-8_13
14. Miyakawa M. Criteria for selecting a variable in the construction of efficient decision trees / M. Miyakawa // IEEE Transactions on Computers. – 1989. – Vol. 38, № 1. – P. 130–141.
15. Tolosi L. Classification with correlated features: unreliability of feature ranking and solutions / L. Tolosi, T. Lengauer // Bioinformatics. – 2011. – Vol. 27, № 14. – P. 1986–1994. DOI:10.1093/bioinformatics/btr300
16. Chaudhuri A. Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York : Chapman & Hall, 2005. – 416 p. DOI: 10.1201/9781420028638
17. Subbotin S.A. Methods of sampling based on exhaustive and evolutionary search / S. A. Subbotin // Automatic Control and Computer Sciences. – 2013. – Vol. 47, No. 3. – P. 113–121. DOI: 10.3103/s0146411613030073
18. Subbotin S.A. The sample properties evaluation for pattern recognition and intelligent diagnosis / S. A. Subbotin // Digital Technologies : 10th International Conference, Zilina, 9–11 July 2014 : proceedings. – Los Alamitos: IEEE, 2014. – P. 332–343. DOI: 10.1109/dt.2014.6868734
19. Lavrakas P.J. Encyclopedia of survey research methods. – Thousand Oaks: Sage Publications, 2008. – Vol. 1–2. – 968 p. DOI: 10.4135/9781412963947.n159
20. Łukasik S. An algorithm for sample and data dimensionality reduction using fast simulated annealing / S. Łukasik, P. Kulczycki // Advanced Data Mining and Applications, Lecture Notes in Computer Science. – Berlin : Springer, 2011. – Vol. 7120. – P. 152–161. DOI: 10.1007/978-3-642-25853-4_12
21. Subbotin S.A. The training set quality measures for neural network learning / S. A. Subbotin // Optical Memory and Neural Networks (Information Optics). – 2010. – Vol. 19, No. 2. – P. 126–139. DOI: 10.3103/s1060992x10020037
22. Cheng Q. Multifractal Modeling and Lacunarity Analysis / Q. Cheng // Mathematical Geology. – 1997. – Vol. 29, No. 7. – P. 919–932. DOI:10.1023/A:1022355723781
23. Eftekhari A. Fractal Dimension of Electrochemical Reactions / A. Eftekhari // Journal of the Electrochemical Society. – 2004. – Vol. 151, No. 9. – P. E291–E296. DOI:10.1149/1.1773583.
24. Evaluating the fractal dimension of profiles / [B. Dubuc, J. Quiniou, C. Roques-Carmes et al.] // Physical Review. – 1989. – Vol. 39, No. 3. – P. 1500–1512. DOI:10.1103/PhysRevA.39.1500
25. Camastra F. Data Dimensionality Estimation Methods: A survey / F. Camastra // Pattern Recognition. – 2003. – Vol. 36, No. 12. – P. 2945–2954. DOI: 10.1016/S0031-3203(03)00176-6
26. A fast and effective method to find correlations among attributes in databases / [P. M. de Sousa, C. Traina, A. J. M. Traina et al.] // Data Mining and Knowledge Discovery. – 2007. – Vol. 14, Issue 3. – P. 367–407. DOI: 10.1007/s10618-006-0056-4
27. Roberts A. Unbiased estimation of multi-fractal dimensions of finite data sets / A. Roberts, A. Cronin // Physica A: Statistical Mechanics and its Applications. – 1996. – Vol. 233, No. 3–4. – P. 867–878. DOI:10.1016/s0378-4371(96)00165-3
28. Kumaraswamy K. Fractal Dimension for Data Mining [Electronic resource] / K. Kumaraswamy. – Access mode: https://www.ml.cmu.edu/research/dap-papers/skkumar_kdd_project.pdf.
29. Li J. An improved box-counting method for image fractal dimension estimation / J. Li, Q. Du, C. Sun // Pattern Recognition. – 2009. – Vol. 42, No. 11. – P. 2460–2469. DOI:10.1016/j.patcog.2009.03.001.
30. Signal attenuation and box-counting fractal analysis of optical coherence tomography images of arterial tissue / [D. P. Popescu, C. Flueraru, Y. Mao et al.] // Biomedical Optics Express. – 2010. – Vol. 1, No. 1. – P. 268–277. DOI:10.1364/boe.1.000268
31. Takens F. On the numerical determination of the dimension of an attractor / F. Takens // Dynamical Systems and Bifurcations Workshop Groningen, 16–20 April 1984 : proceedings. – Berlin: Springer, 1985. – P. 99–106. DOI: 10.1007/bfb0075637
32. Субботин С. А. Метрики качества выборок данных и моделей зависимостей, основанные на фрактальной размерности / С. А. Субботин // Радіоелектроніка, інформатика, управління. – 2017. – № 2. – С. 70–81.
33. Zong-Chang Y. Establishing structure for artificial neural networks based-on fractal / Y. Zong-Chang // Journal of Theoretical and Applied Information Technology. – 2013. – Vol. 49, No. 1. – P. 342–347.
34. Crişan D.A. Fractal dimension spectrum as an indicator for training neural networks / D. A. Crişan, R. Dobrescu, // Universitatea Politehnica Bucuresti Sci. Bull. Series C. – 2007. – Vol. 69, No. 1. – P. 23–32.
35. Subbotin S. A. The neural network model synthesis based on the fractal analysis / S. A. Subbotin // Optical Memory and Neural Networks. – 2017. – Vol. 26, No. 4. – P. 257–273.
36. Fisher Iris dataset [Electronic resource].– Access mode: https://archive.ics.uci.edu/ml/datasets/Iris
37. The plant recognition on remote sensing results by the feed-forward neural networks / [V. Dubrovin, S. Subbotin, S. Morshchavka, D. Piza] // International Journal of Smart Engineering System Design. – 2001. – Vol. 3, No. 4. – P. 251–256.
38. Arrhythmia Data Set [Electronic resource]. – Access mode: http://archive.ics.uci.edu/ml/datasets/arrhythmia

DOI 10.15588/1607-3274-2020-1-11

e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2020. № 1
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2020. № 1

ПРОГРЕСИВНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

PROGRESSIVE INFORMATION TECHNOLOGIES

ПРОГРЕССИВНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

UDC 004.9

МЕТОД АВТОРИФІКАЦІЇ ТЕКСТУ НАУКОВО-ТЕХНІЧНИХ ПУБЛІКАЦІЙ НА ОСНОВІ ЛІНГВІСТИЧНОГО АНАЛІЗУ КОЕФІЦІЄНТІВ МОВНОЇ РІЗНОМАНІТНОСТІ

Висоцька В. А. – канд. техн. наук, доцент, доцент кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. Авторифікація авторства тексту є технікою визначення автора тексту, коли неоднозначно, хто її написав. Це корисно, коли декілька людей претендують на авторство однієї публікації або у випадках, коли ніхто не претендує на авторство текстового контенту, наприклад, так звані тролі в соціальних мережах під час інформаційної війни. Складність проблеми авторського тексту, очевидно, експоненціально вища, більша кількість вірогідних авторів. Наявність авторських текстових зразків також є суттєвою при просуненні цієї проблеми. Атрибуція авторського тексту включає наступні три проблеми:

– виявлення автора текстового автора з групи імовірних або очікуваних авторів, де автор завжди знаходиться у групі підозрюваних;

– не ідентифікація автора текстового автора з групи вірогідних або очікуваних авторів, де автор може не бути в групі підозрюваних;

– оцінка можливості даного тексту, написаного даним автором чи ні.

Тому задача автоматичного визначення автора текстового контенту науково-технічного спрямування є актуальною й потребує нових (досконаліших) підходів до її розв'язування.

Метою дослідження є розроблення методу визначення автора у україномовних текстах на основі технології лінгвометрії.

Метод. Розроблено лінгвометричний метод алгоритмічного забезпечення процесів контент-моніторінгу для розв'язання задачі автоматичного визначення автора україномовного текстового контенту на основі технології статистичного аналізу коефіцієнтів мовної різноманітності. Проведено декомпозицію методу визначення автора на основі аналізу таких коефіцієнтів мовлення як лексична різноманітність, ступінь (міра) синтаксичної складності, зв'язність мовлення, індекси винятковості та концентрації тексту. Проаналізовані також параметри авторського стилю як кількість слів у певному тексті, загальна кількість слів цього тексту, кількість речень, кількість прийменників, кількість сполучників, кількість слів із частотою 1, та кількість слів із частотою 10 та більше. Особливостями розробленого є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів. Тобто при аналізі лінгвістичних одиниць типу слів, враховувалась належність до частини мови та відмінювання в межах цієї частини мови. Для цього провадився аналіз флексій цих слів для класифікації, виділення основи для формування відповідних алфавітно-частотних словників. Наповнення цих словників в подальшому враховувалися на наступних кроках визначення авторства тексту як розрахунок параметрів та коефіцієнтів авторського мовлення. Для індивідуального стилю письменника показовими є саме службові (стопові або опорні) слова, оскільки вони ніяк не пов'язані з темою і змістом публікації.

Результати. Проведено порівняння результатів на множині 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу.

Висновки. Виявлено, що для обраної експериментальної бази з понад 200 робіт найкращих результатів за критерієм щільності досягає метод аналізу статті без початкової обов'язкової інформації як анотації та ключові слова різними мовами, а також списку літератури.

КЛЮЧОВІ СЛОВА: текстовий контент, NLP, контент-моніторінг, стоп-слова, контент-аналіз, статистийний лінгвістичний аналіз, квантитативна лінгвістика.

АБРЕВІАТУРА

ІС – інформаційна система;

ІТ – інформаційна технологія.

DOI 10.15588/1607-3274-2020-1-12

НОМЕНКЛАТУРА

S – система визначення автора;

A – множина авторських статтей;

D – множина статтей для дослідження;

K – множина коефіцієнтів авторського мовлення;

H – множина коефіцієнтів мовлення невідомого автора досліджуваного тексту;

L – словник службових слів;

C – результати порівняння коефіцієнтів мовлення відомих авторів та досліджуванного тексту;

α – оператор парсингу тексту для визначення множини параметрів авторського мовлення;

μ – оператор визначення коефіцієнтів авторського мовлення відомих/досліджуваних публікацій;

ψ – оператор порівняння коефіцієнтів авторського мовлення відомих/досліджуваних публікацій;

χ – оператор формування множини публікацій з подібними значеннями коефіцієнтів авторського мовлення;

ω – оператор машинного навчання системи на основі попередньо зібраної статистики розрахунків коефіцієнтів авторського мовлення;

θ – оператор розрахунку ступення належності досліджуваного тексту конкретному автору з множини потенційних авторів;

K_r – коефіцієнт розповсюдженості;

N_p – відношення кількості підвибірок з певною лінгвістичною одиницею;

N_z – загальна кількість підвибірок;

K_l – коефіцієнт лексичної різноманітності;

W – кількість слів у певному тексті;

N – загальна кількість слів цього тексту;

K_s – коефіцієнт синтаксичної складності;

P – кількість окремих речень;

W – кількість слів у всьому тексті;

K_z – коефіцієнт зв'язності мовлення;

Z – кількість прийменників;

S – кількість сполучників;

I_{wt} – індекс винятковості тексту;

W_1 – кількість слів із частотою 1;

I_{kt} – індекс концентрації тексту;

W_{10} – кількість слів із частотою 10 та більше;

F – частота слова в частотному словнику;

i – ранг слова в частотному словнику;

k – довжина слова у фонемах;

C – стала;

r – ранг слова у фонемах;

m – кількість значень слова;

f – частота слова;

y – середня довжина складових;

x – довжина мовної конструкції;

b – показник, що характеризує динаміку зміни довжини складників (закон діє, якщо $b<0$);

p_x – ймовірність використання слова, яке має x значень;

ω – середня кількість значень слова у словнику.

ВСТУП

Важливими завданнями мовознавства на основі лінгвометрії є створення і порівняння словників (у тому числі частотних та статистичних), автоматичних словників, тезаурусів, систем стенографії, автоматичне визначення мови, інформаційний пошук тощо [1]. Наприклад, для моделювання процесів інформаційного пошуку знаходять статистичні і перехідні ймовірності морфем тексту [2]. На основі побудованих таблиць моделюють перевірку досліджуваного слова на наявність помилки, пропонують кілька найбільш ймовірних варіантів [3]. Лінгвометрія – галузь прикладної лінгвістики, що виявляє, вимірює та аналізує кількісні характеристики одиниць різних рівнів мови чи мовлення. Одним зі способів охарактеризувати літературне багатство тексту є оцінювання характеру використання мовних одиниць на всіх мовних рівнях. Це дає змогу ототожнювати поняття багатство і різноманітність мовлення. У свою чергу стилеметрія як підрозділ прикладної лінгвістики виявляє та аналізує кількісні характеристики певного функціонального стилю мови чи мовлення авторів текстового контенту, тобто авторської атрибуції [4]. Атрибуція полягає у визначенні методома квантитативної лінгвістики достовірності, автентичності авторського твору, його автора, місця й часу створення на основі аналізу технологічних і стилістичних закономірностей та особливостей коефіцієнтів мовної різноманітності конкретного авторач і/або конкрентного текстового твору [5]. Наприклад, однією із відомих мовознавчих проблем є процес визначення авторської атрибуції уривків певного текстового контенту [6]. Для цього обчислють частоти слововживань у запропонованих уривках [7]. Використовуючи частотні словники авторської творчості загалом чи окремих його творів, визначають автора твору (або твір – якщо це дозволяє словник) [8]. Недоліком є збереження або автоматичне генерування великих масивів даних у вигляді частотних словників авторських творів [9]. Опрацювання таких словників вимагає багато часу, а збереження – багато ресурсів [10]. У свою чергу, є автори з малочисельною творчістю, що унеможливлю точне відворення результатів аналізу авторської атрибуції [11]. Відомий метод датування для визначення тривалості роздільного існування двох споріднених мов, ґрунтується на припущенні про те, що основна частина лексичного складу будь-якої мови (ядерна лексика) змінюється з однаковою швидкістю і вимагає підрахунку процентного співвідношення спільних елементів у основному словнику [12]. Модифіковані методи глотохронології застосовують для визначення динаміки зміни авторського мовлення в його текстовому контенті на протязі тривалого часу для датування наближеного періоду, в якому був створений конкртниий текст твору цього автора [13]. Тому задача автоматичного визначення автора текстового контенту є актуальною й потребує нових (досконаліших) підходів до її

DOI 10.15588/1607-3274-2020-1-12

розв'язування, наприклад, на основі статистичного аналізу коефіцієнтів мовної різноманітності [14].

Метою дослідження є розроблення методу визначення автора в україномовних текстах на основі технології лінгвометрії. Для досягнення мети були поставлені такі завдання:

– на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному тексті розробити алгоритми визначення автора тексту;

– розробити програмне забезпечення контент-моніторінгу для визначення автора в україномовних текстах на основі лінгвометричного аналізу визначенних стопових слів текстового контенту;

– здійснити аналіз результатів експериментальної апробації запропонованого методу контент-моніторінгу для визначення автора в україномовних наукових текстах технічного профілю.

1 ПОСТАНОВКА ПРОБЛЕМИ

Систему визначення автора в україномовних текстах на основі технології лінгвометрії подамо як кортеж $S=<A, D, L, K, H, C, \alpha, \mu, \psi, \chi, \theta, \omega>$. Вагомим значенням у квантитативній лінгвостатистиці є розподіл (дистрибуція) лінгвістичної одиниці у тексті – присутність лінгвістичної одиниці в різних (зазвичай рівних) підвибірках (уривках) [15]. Якщо досліджувана лінгвістична одиниця функціонує тільки в одній нідвибірці, хоча й з високою частотою, то така вибірка є нерепрезентативною стосовно цієї лінгвістичної одиниці [16]. Важливо, коли досліджувана лінгвістична одиниця є рівномірно розподіленою в генеральній сукупності на прикладі науково-технічни статтей україномовних публікацій [17]. Відповідно визначають критерії для розрахунку коефіцієнтів авторського мовлення як $K=\mu(L)°\alpha(A)$ або $H=\mu(L)°\alpha(D)$, де $K=\{ K_r, K_l, K_s, K_z, I_{wt}, I_{kt}\}$ та $H=\{K_r, K_l, K_s, K_z, I_{wt}, I_{kt}\}$ відповідно авторського відомого тексту та досліджуваного тексту статті відповідно. Для цього аналізують коефіцієнт розповсюдженості [18]: $K_r=N_p/N_z$.

Проте характеристики, одержані на матеріалі вибірки, зазвичай відрізнятються від реальних характеристик генеральної сукупності, оскільки завжди присутня в квантитативній лінгвостатистиці відносна неточність дослідження [19]. Розподіл частоти лінгвістичних одиниць мови в текстовому контенті має певну регулярність і утворює його статистичну (частотну, ймовірнісну) структуру [20]. Такий розподіл є відмінним для кожної з мовних елементів – лексем, морфем, фонем тощо [21]. Тому лінгвостатистичні параметри авторських стилів, встановлені на різних рівнях (фонемних, морфемних, *N*-грамних, лексемних тощо), мають неоднакову стилеідентифікованму потужність автоського мовлення для різних пар стилів [22]. Наприклад, споріднені стилі чіткіше розмежовані на синтаксичному рівні, а менш споріднені – на лексичному [23]. Для цього автоматично створюють частотні словники певних лінгвістичних одиниць та завдяки ним аналізують середню повторюваність слова в тексті, коефіцієнт hapax legomena (слова, які мають частоту 1 у досліджуваній вибірці), індекс винятковості, індекс концентрації тощо [1–5, 14, 24].

Розрахунок коефіцієнтів мовної різноманітності повинен припускати взаємозв'язок таких коефіцієнтів, як:

– лексична різноманітність ($K_l=W/N$): відношення кількості слів до загальної кількості словоформ тексту, значення коефіцієнта лежить у межах [0;1];

– ступінь (міра) синтаксичної складності ($K_s=1–P/W$): відношення кількості речень до кількості слів певного тексту [14];

– зв'язність мовлення ($K_z=(Z+S)/(3P)$): відношення кількості прийменників і сполучників до кількості окремих речень;

– індекс винятковості тексту ($I_{wt}=W_1/W$): варіативність лексики, тобто частка тексту, яку займають слова, що трапилися 1 раз;

– індекс концентрації тексту ($I_{kt}=W_{10}/W$): частка тексту, яку займають слова, що трапилися 10 разів і більше [14].

Оскільки коефіцієнт – величина абсолютна, можна у певних межах нехтувати довжиною порівнюваних текстів [46]. Теоретичний інтерес складає дослідження внутрішньої «динаміки» тексту в частині співставлення коефіцієнтів з різних його ділянок між собою та із загальним для всього тексту коефіцієнтом [1–5, 14, 41, 47]:

– для лексичної різноманітності чим більшим є отримуваний десятковий дріб, тим вищою є лексична різноманітність досліджуваного тексту [14];

– для синтаксичної складності чим більшим є дріб (в межах [0; 1]), тим багатослівнішими загалом є речення такого тексту, а отже, – вища можливість різноманітності синтаксичних відношень між словами в окремому реченні [14];

– для зв'язності мовлення дорівнює одиниці, коли в одному реченні є три сполучні елементи (прийменники і сполучники) [14].

Далі необхідно розрахувати значення ступенів приналежності досліджуваного тексту відповідним авторам із списку подібних за значеннями коефіцієнтами мовлення в допустимих межах:

$$C=\omega(K,H)°\theta(K,H)°\chi(K,H)°\psi(K,H).$$

Результатом функціонування системи буде рангований список потенційних аторів досліджуваного тексту україномовної науково-технічної публікації. Зменшення позицій тексту суттєво залежить від кількості публікації авторів, часового проміжку самих публікації, наявності достовірних даних про приналежність тексту конкретному автору, обсягу статистичних даних для машинного навчання системи для формування частотних словників використання сд\лужбових слів конкретним автором.

DOI 10.15588/1607-3274-2020-1-12

2 АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

За даними ЧС обчислюють такі характеристики як багатство словника, індекс різноманітності (K_l) – відношення обсягу словника лексем (W) до обсягу тексту (N), тобто $K_l=W/N$. Згідно табл. 1 найрізноманітніша, найбагатша лексика – у поезії, далі за спадом – у художній прозі, розмовно-побутовому стилі, публіцистиці, науковому та офіційно-діловому стилі [14, 25].

Середня повторюваність слова у тексті A є відношенням обсягу тексту N до обсягу словника лексем W (обернена до індексу різноманітності), тобто $A=N/W$ [26]. За даними ЧС, кожне слово у розмовно-побутовому стилі в середньому вжито 14 разів, а в науковому стилі – 17 [27].

Таблиця 1 – Результати коефіцієнтів мовлення згідно стилів української мови [14]

Стиль	W/N	W_1/N	W_1/W	W_{10}/W	W_{10t}/N
науковий	0,059	0,427	0,025	0,189	0,890
публіцистичний	0,070	0,450	0,031	0,121	0,804
діловий	0,030	0,280	0,0085	0,303	0,935
поетичний	0,103	0,495	0,052	0,098	0,789
художньої прози	0,067	0,430	0,029	0,149	0,821
розмовний	0,073	0,465	0,034	0,161	0,789

Індекс винятковості характеризує варіативність лексики, тобто частку тексту (словника), яку займають слова, що трапилися 1 раз (табл. 1) [28]:

– словника I_{wt} – відношення кількості лексем із частотою 1 W_1 до загальної кількості лексем: $I_{wt}=W_1/W$ [14];

– тексту I_t відношення кількості лексем із частотою 1 W_1 до обсягу тексту N: $I_t=W_1/N$ [14].

Індекс концентрації вказує на частку тексту (словника), яку займають слова, що трапилися 10 разів і більше (табл. 1) [29]:

– словника I_{kt} – відношення кількості слів у словнику з абсолютною частотою 10 і більше (W_{10}) до загальної кількості слів у словнику (W): $I_{kt}=W_{10}/W$ [14];

– тексту I_{tn} – відношення суми абсолютних частот слів з абсолютною частотою 10 і більше W_{10t} до обсягу тексту N: $I_{tn}=W_{10t}/N$ [14].

Мовлення надає перевагу невеликій кількості одиниць, які часто використовують [30]. Формують ядро будь-якої мовленнєвої підсистеми, тоді як переважна кількість одиниць є низькочастотними [31]. Цю закономірність зауважив ще учений Дьюї на поч. XX ст., назвавши її законом переваги [32]. Детальніше дослідив цю закономірність німецький мовознавець Дж. Ціпф, сформулювавши закон Zipf's law, який встановлює залежності [33]:

– частоти слова та його рангу у словнику: у частотнішого слово вищий ранг при $F\cdot i$=const [34];

– частоти слова та його довжини: чим частотніше слово, тим воно коротше при $k=C\lg r$ [35];

– частоти слова та кількості його значень: чим частотніше слово, тим воно багатозначніше при $m=C\sqrt{f}$ [36];

– частоти слова та його походження: чим давніше слово, тим воно частотніше [37].

Згідно закону німецького мовознавця П. Менцерата довжина мовної конструкції (слова, словосполучення, надфразової єдності, речення) обернено пропорційна до довжини її складових (складів, слів, словосполучень і т. д.), тобто чим довша мовна конструкція, тим коротші її складові [14]. Згідно досліджень Ґ. Альтманна $y=ax^b$.

Закон Крилова встановлює залежність між кількістю багатозначних слів та частотою [14]:

$$p_x=1/2^x,\quad px=(\omega-1)^{x-1}/\omega^x.$$

Деякі основні кількісні характеристики мови дуже прості. Наприклад, різниця між кількістю слів (10^4–10^5), кількістю морфем (декілька тисяч), кількістю складів (від декількох сотень до декількох тисяч) і кількістю фонем (від 10 до 80) [31–37]. Висловлюють припущення, що такі співвідношення пов'язані із властивістю людської пам'яті. Зазначимо також, що чим частотніше слово, тим швидше людина його зможе пригадати. Однак відсутні дослідження в галузі залежності змін коефіцієнтів лексичного авторського мовлення на протязі періоду його творчості.

3 МАТЕРІАЛИ ТА МЕТОДИ

Виявлено [14], що текст україномовної казки має K_z=0,77, а текст україномовної наукової статті – 3,0, тобто зв'язність у другому тексті у 3,9 разів сильніша, ніж у першому. Офіційних стандартів для коефіцієнтів різноманітності мовлення для K_l та K_s не існує, але орієнтиром для співставлення та оцінювання якогось тексту в однорідній групі текстів є середньостатистична норма величини коефіцієнта для рівних за довжиною уривків. Мінімальним розміром (довжиною) уривка приймемо 100 слів, вважатимемо, що коефіцієнти тут уже стабілізуються, відображаючи реальні особливості мови автора. Близькість або віддаленість окремого індивідуального коефіцієнта від середнього служить основою для оцінювання різноманітності мовлення у відповідному тексті. Задовільними вважаються тексти, коефіцієнти різноманітності яких потрапляють у зону середніх квадратичних відхилень D від певного середнього

$$D=\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{n}-(\overline{x})^2}\ \text{[14]}.$$

Аналіз та інтерпретація на лінгвістичному рівні стилістичних особливостей і закономірностей письменницького стилю певного автора (або певної літературної епохи) включає подані найосновніші етапи, подані в алг. 1 [14].

Алгоритм 1. Аналіз та інтерпретація на лінгвістичному рівні стилістичних особливостей і

DOI 10.15588/1607-3274-2020-1-12

закономірностей письменницького стилю певного автора

Етап 1. Відбір та первинне опрацювання текстового контенту. Для відбору будують фільтри тексту за параметрами (основна мова тексту, обсяг текстової вибірки, часовий проміжок публікації, джерео публікації, формат тощо). Основними кроками первинного опрацювання тексту є:

– приведення його до єдиного формату (наприклад усунення тегів, якщо попередня публікація є у Інтернет-ресурсі у виглдяі статичної сторінки);

– усунення інформаційного шуму (рисунків, формул, список літеру тари, анотації іншими мовами тощо), який не впливає на результат, але збільшує час опрацювання;

– приведення до єдиного обсягу (скорочення у разі потреби, забираючи неінформативні ділянки початку та закінчення тексту).

Етап 2. Лематизація текстових лінгвістичних одиниць. Об'єднання словоформ під лемою мови [14].

Етап 3. Усунення неоднорідності текстових лінгвістичних одиниць. Розв'язання проблеми неоднорідності текстових лінгвістичних одиниць, наприклад, із погляду відношення до різних видів мови (авторська, не авторська і т. п.).

Етап 4. Побудова системи частотних словників, організація на основі статистичних розподілів у потрібних частотних шкалах. Частотний словник – тип словника, де наведено кількість вживань (частоту) певної лінгвістичної одиниці мови (складу, слова, словоформи, словосполучення, ідіоми, фразеологізму) в різних текстах певного обсягу. Зазвичай, подають абсолютну та відносну частоту вживання мовних одиниць, словникові статті розміщують за спаданням частот.

Етап 5. Пошук параметрів, що адекватно відображають структуру частотного словника. Такі параметри дають змогу сформулювати кілька основних лінгвостатистичних методів дослідження тексту:

– метод опорних слів (підрахунок загальної частоти вживання та знаходження відсоткового складу службових слів: прийменників, сполучників, часток);

– метод розділових знаків (підрахунок лише кількості внутрішніх і зовнішніх розділових знаків);

– метод слів (підрахунок лише слів певної довжини);

– метод речень (підрахунок лише речень визначеної довжини);

– синтаксичний метод (підрахунок розділових знаків, слів і речень певної довжини);

– комбінований (поєднання синтаксичного методу і опорних слів).

Етап 6. Перевірка параметрів на ефективність. Аналіз та порівння отримани результатів на відомих авторських творах для визначення закономірностей впливу авторської стилістики на формування авторської структури частотного словника за цими параметрами.

Етап 7. Математичне моделювання лексикостатистичних розподілів.

Етап 8. Побудова статистичних класифікацій, тобто авторських еталонів, що відображають стилістичні закономірності в межах творів певного автора чи певної літературного стилю з врахуванням літературної епохи та особливостей мови, на якій написані самі аналізовані твори.

Етап 9. Інтерпретація результатів із позицій стилістичних уявлень у визначеному часовому проміжку, загальної й авторської стилістики з врахуванням часових параметрів. Таким чином також вирішемо завдання авторської атрибуції, яке сформулюємо наступним чином. Нехай існує статистично опрацьований доробок автора (еталон). Необхідно оцінити належність певних уривків до еталону із застосуванням відповідних методів. Графічне зображення відносної частоти появи службових слів в Уривку 4 та в еталоні подане на рис. 1. Коефіцієнт кореляції для службових слів у цьому випадку складає R_{e-U4}=0,7326. Наведемо також коефіцієнти кореляції для кожного зі службових слів для уривків 1–4 (табл. 4). Аналізуючи коефіцієнти кореляції для службових слів, приходимо до висновку, що ймовірність належності уривків до досліджуваного еталону найбільшою є для Уривку 4, за ним – Уривок 2, Уривок 1, Уривок 3. Зауважимо, що для всіх чотирьох уривків простежуються стабільно високі коефіцієнти кореляції для часток, що можемо розуміти як відсутність впливу часток на авторський стиль. Додатково для уривків проаналізуємо частотності поав лише прийменників і сполучників, знайдемо відповідні коефіцієнти кореляції та порівняємо результати (табл. 2).

Рисунок 1 – Відносна частота появи службових слів в Уривку 4 та в еталоні

DOI 10.15588/1607-3274-2020-1-12

Таблиця 2 – Коефіцієнти кореляції для службової частини мови та кожного з уривків

№	Прийменник	Сполучник	Частка	R_{e-U}	R'_{e-U}
1	0,72	0,79	1	0,6076	0,6900
2	0,4928	0,5714	0,9580	0,7066	0,4913
3	0,1517	0,1624	0,8800	0,2810	0,2254
4	0,5639	0,9544	0,9594	0,7326	0,6905

Уривок 4 так і залишився найімовірнішим кандидатом щодо належності його до еталону, а наступним із незначним відривом став Уривок 1, далі – Уривок 2. Уривок 3, як і у попередньому дослідженні, має найменшу ймовірність належати до еталону. Для підтвердження результатів звернемося до [1–4], з яких узято уривки для дослідження.

4 ЕКСПЕРИМЕНТИ

Під час дослідження розроблено систему з можливістю обрання мови/мов аналізованого контенту, реалізована на Web-ресурсі Victana (рис. 2). Аналізуючи складові формул для оцінки багатства твору, приходимо до висновку, що треба знайти такі величини як кількість слів і словоформ, речень, сполучників і прийменників, слів із частотою 1 та не меншою за 10. На сервері після запуску процесу розрахунку коефіцієнтів різноманітності тексту запускається алгоритм аналізу цього тексту (алг. 2).

Алгоритм 2. Аналізу стилю автоського мовлення.

Етап 1. Перевірка довжини тексту – лишнє відсікається.

Етап 2. Очищення досліджуваного тексту (цифри, спецсимволи, формули, рисунки).

Етап 3. Визначення кількості речень P.

Етап 4. Визначення кількості слів у тексті N.

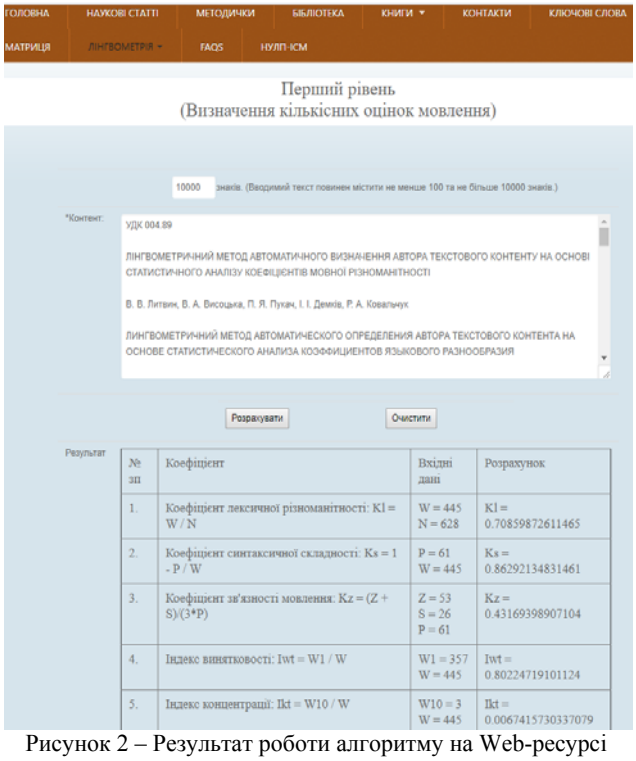


Рисунок 2 – Результат роботи алгоритму на Web-ресурсі Victana (http://victana.lviv.ua/nlp/linhvometriia)

DOI 10.15588/1607-3274-2020-1-12

Етап 5. Визначення кількості слів W (за частотним словником основ слів).

Етап 6. Розрахунок коефіцієнта лексичної різноманітності: $K_l=W/N$.

Етап 7. Розрахунок коефіцієнта синтаксичної складності: $K_s=1–P/W$.

Етап 8. Визначення кількості слів, що зустрілися точно один раз, тобто W_1.

Етап 9. Розрахунок індексу винятковості тексту : $I_{wt}=W_1/W$.

Етап 10. Визначення кількості слів, що зустрілися більше 9 разів, тобто W_{10}.

Етап 11. Розрахунок індексу концентрації тексту: $I_{kt}=W_{10}/W$.

Етап 12. Визначення кількості прийменників Z.

Етап 13. Визначення кількості сполучників S.

Етап 14. Розрахунок коефіцієнта зв'язності мовлення: $K_z=(Z+S)/(3\cdot P)$.

Етап 15. Виведення результатів на ресурсі Victana.

Аналізуючи складові формул для оцінення багатства твору, бачимо, що треба знайти кількість речень, слів і словоформ, прийменників і сполучників, слів із частотою 1 та частотою, не меншою за 10. Для зручності внесемо знайдені дані у таблицю. На інформаційному ресурсі передається сформована таблиця (табл. 3) та отримані результати дослідження виводяться на екран. Спираючись на викладене вище, оцінимо багатство уривків творів одноосібних наукових статей технічного спрямування Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі» за період 2001–2017 рр. за допомогою коефіцієнтів різноманітності та зв'язності мовлення, індексів винятковості та концентрації тексту. Для аналізу виберемо частину першу (10000 знаків) кожної статті (алг. 3).

Алгоритм 3. Аналіз статистики функціонування системи виявлення множини стопових слів із 215 наукових статей технічного спрямування

Етап 1. Аналіз 100 наукових статей на визначення діапазону оптимального розміру досліджуваного тексту. Спочатку були проаналізовані тексти в повному обсязі, а потім ці тексти були проаналізовані на різні величини знаків. Результати показали, що

Таблиця 3 – Приклад результату роботи алгоритму аналізу стилю автора публікації на ресурсі Victana

Коефіцієнт	Дані	Розрахунок
лексичної різноманітності: $K_l=W/N$	W=184; N=295	K_l=0,6237
синтаксичної складності: $K_s=1–P/W$	P=18; W=184	K_s=0,902
зв'язності мовлення: $K_z=(Z+S)/(3\cdot P)$	Z=20; S=28; P=18	K_z=0,889
винятковості: $I_{wt}=W_1/W$	W_1=141; W=184	I_{wt}=0,7663
концентрації: $I_{kt}=W_{10}/W$	W_{10}=2; W=184	I_{kt}=0,01

оптимальним дослідженням текстів є діапазон [100;10000] знаків. Менше 100 знаків – неінформативна отримана інформація, часто значення коефіцієнтів різних авторів подібні, а одного ж автора на різних тестах – суттєво різняться. Якщо більше 10000 знаків – суттєво коефіцієнти вже не змінюються, але аналоги для дослідження мають різну довжину і з-за браку різноманітності аналогів великої довжини, було обрано максимальне число для аналізу 10000.

Етап 2. Аналіз понад 200 одноосібних робіт технічного спрямування понад 50 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу.

Етап 3. Аналіз понад 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу.

Етап 4. Аналіз понад 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення стилів мовлення цих авторів.

Етап 5. Аналіз отриманих коефіцієнтів мовлення біля 100 різних авторів за період 2001–2017 рр. для визначення підмножини авторів з подібним стилем, що і 4 еталоні роботи (колективні роботи, автори яких присутні серед досліджуваних одноосібних робіт).

Етап 6. Аналіз отриманих результатів на етапі 5. Перевірити, чи в отриманих підмножинах присутні справжні автори цих еталонних текстів. Обрати найкращий алгоритм для визначення стилю автора в україномовних науково-технічних текстах на основі технології квантитативної лінвостатистики

5 РЕЗУЛЬТАТИ

Для чистоти дослідження необхідно проаналізувати, чи впливає час публікації робіт на коефіцієнти різноманітності тексту, тобто чи не змінюються ці коефіцієнти з часом на вибірці тих самих авторів та текстів. Спочатку проаналізуємо як змінюється загальний обсяг слів в однакових за розміром уривках в діапазоні 2001–2017 рр. Як бачимо з часом ті ж самі автори частіше вживають коротші слова (рис. 3а).

З часом коефіцієнт лексичної різноманітності K_l суттєво не змінюється (рис. 3б–3г). Аналогічно з часом коефіцієнт синтаксичної складності K_s також суттєво не змінюється. А ось коефіцієнт зв'язності мовлення K_z з часом за 16 років зменшується, хоча не суттєво. На початках (2001 р.) коливається в діапазоні [0,5;1,2], а в в кінці періоду – в діапазоні [0,4; 0,9] (рис. 4).

Аналогічно порівняємо розподіли індексів винятковості та концентрації (рис. 5). Якщо розмах розподілу суттєво не змінюється в часі для I_{wt}, то для I_{kt} є фіксовані значні зміни. З часом автор цих робіт все частіше повторюють деякі терміни в своїх роботах понад 10 разів, звужуючи коло своїх досліджень. На рис. 5г поданий результат аналізу коефіцієнтів мовлення для однакових за розміром уривках в діапазоні 2001–2017 рр. як мінімальне, максимальне та середнє значення за цей період (визначення коливання значень в цьому часовому проміжку). Більш суттєве коливання спостерігаємо за K_z (рис. 6).

Окремо проаналізуємо розподіл використання всіх словоформ (рис. 6г), слів по одному разу, слів понад 10 разів, вжитих в досліджуваних текстах для однакових за розміром уривках в діапазоні 2001–2017 рр. (рис. 7а). На рис. 7б поданий аналіз вживання прийменників, сполучників та окремих речень в досліджуваних текстах для однакових за розміром уривках в діапазоні 2001–2017 рр., де Z – кількість прийменників, S – кількість сполучників, P – кількість окремих речень. Згідно рис. 7в з часом автори вживають коротші речення для опису предметної області, ніж на початках досліджуваного періоду. Якщо кількість прийменників зменшується, то розподіл вживання сполучників суттєво не зменшується (рис. 7г). На рис. 8а–8б поданий аналіз зміни динаміки вживання слів в досліджуваних текстах за визначений період. На рис. 8в–9г поданий результат аналізу зміни динаміки вживання прийменників, сполучників та речень в досліджуваних текстах за визначений період.

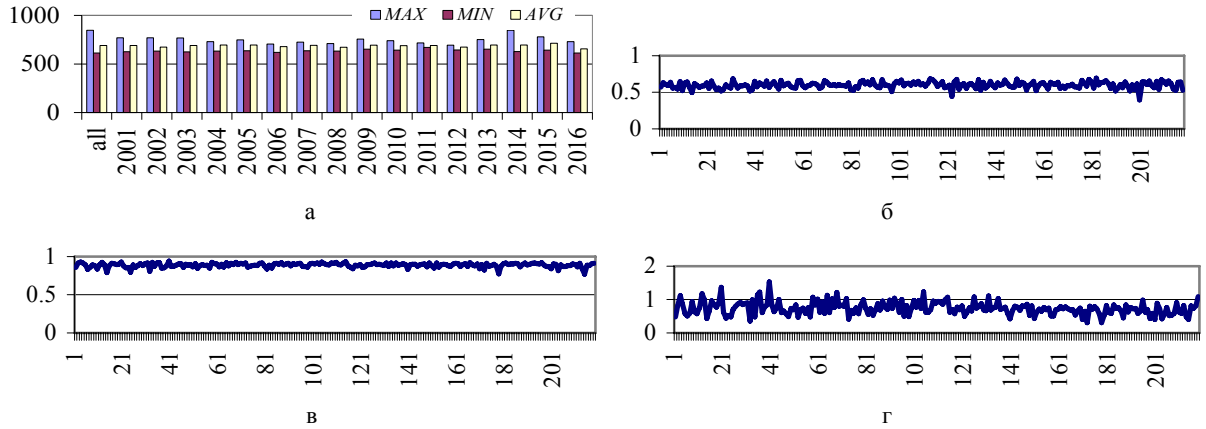


Рисунок 3 – Розподіл: а – слів та коефіцієнтів мовлення для однакових за розміром уривках в діапазоні 2001–2017 рр.: б – K_l; в – K_s; г – K_z

DOI 10.15588/1607-3274-2020-1-12

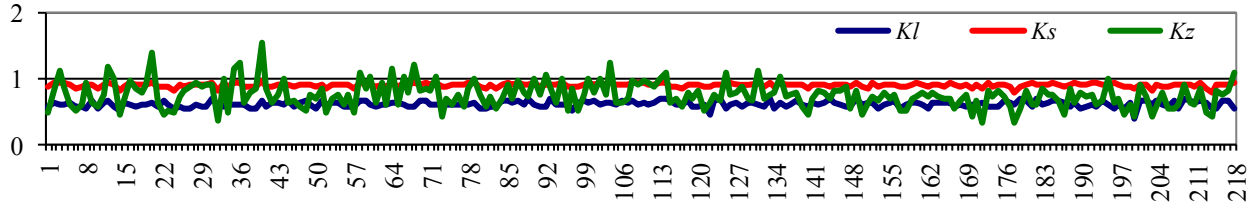


Рисунок 4 – Порівняння розподілу коефіцієнтів мовлення K_l, K_s та K_z

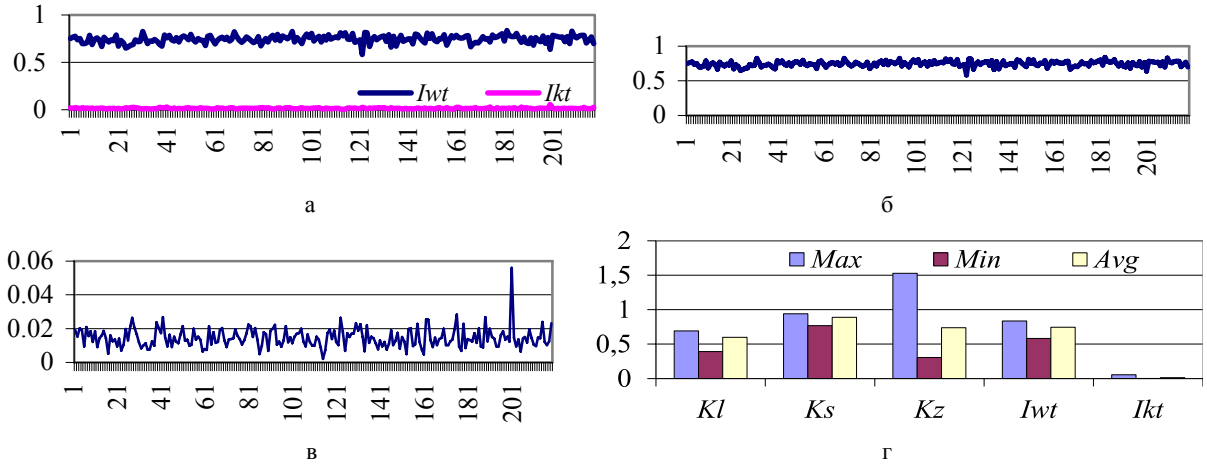


Рисунок 5 – Розподіл індексів мовлення для: а – обох індексів; б – I_{wt}; в – I_{kt}; г – мінімальне, максимальне та середнє значення для всіх коефіцієнтів

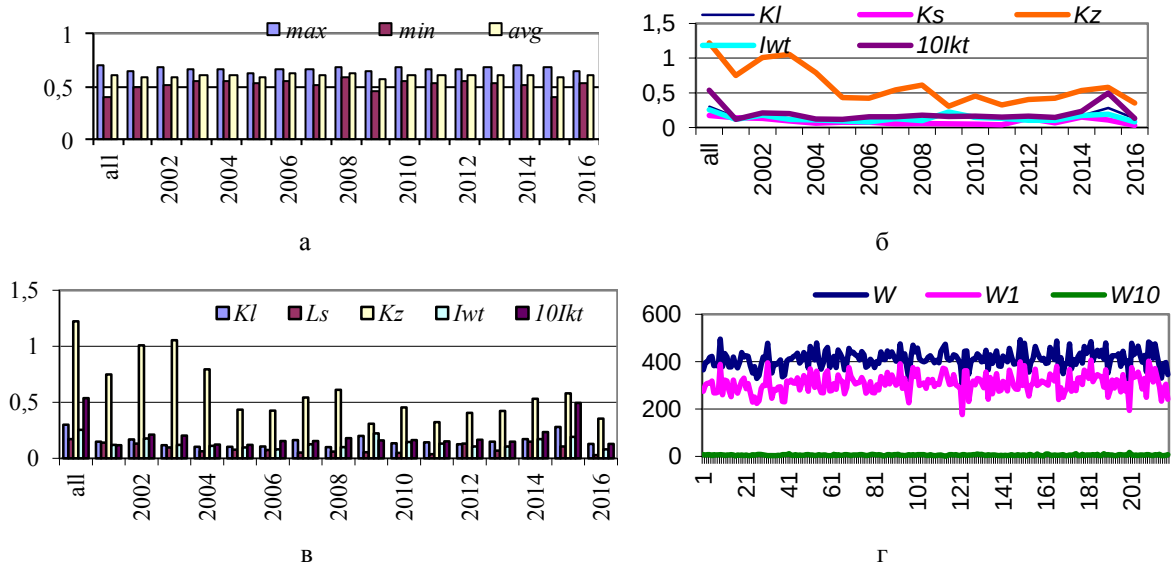


Рисунок 6 – Результат аналізу коефіцієнтів мовлення для однакових за розміром уривках в діапазоні 2001–2017 рр.: а – мінімальне, максимальне та середнє значення за цей період для K_l; б – графік динаміки зміни коефіцієнтів за визначений період; в – гістограма динаміки зміни всі коефіцієнтів за визначений період; г – вживання словоформ (всіх, по 1 разу та понад 10 разів)

DOI 10.15588/1607-3274-2020-1-12

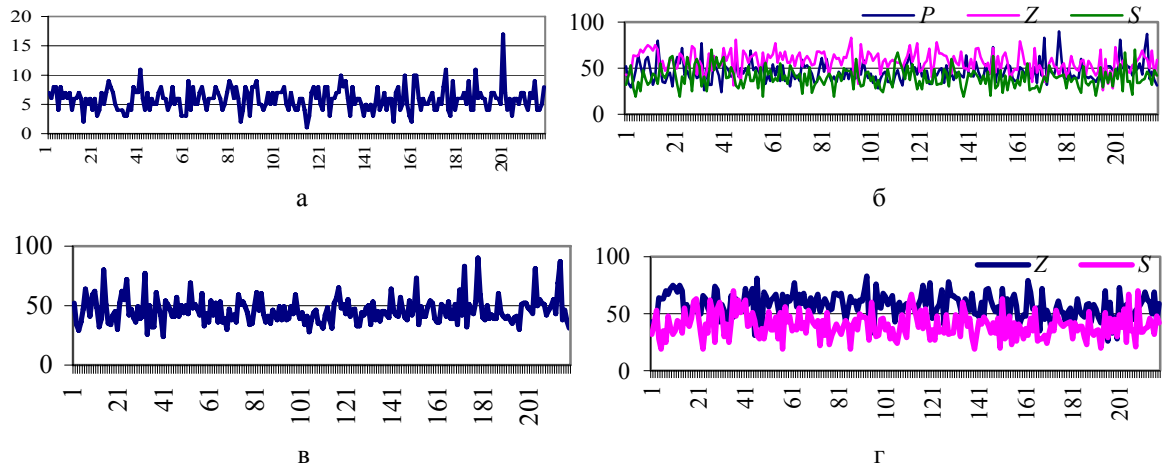


Рисунок 7 – Аналіз частоти вживання слів: *а* – понад 9 разів (W_{10}); *б* – параметрів зв'язності мовлення; *в* – речень; *г* – прийменників, та сполучників

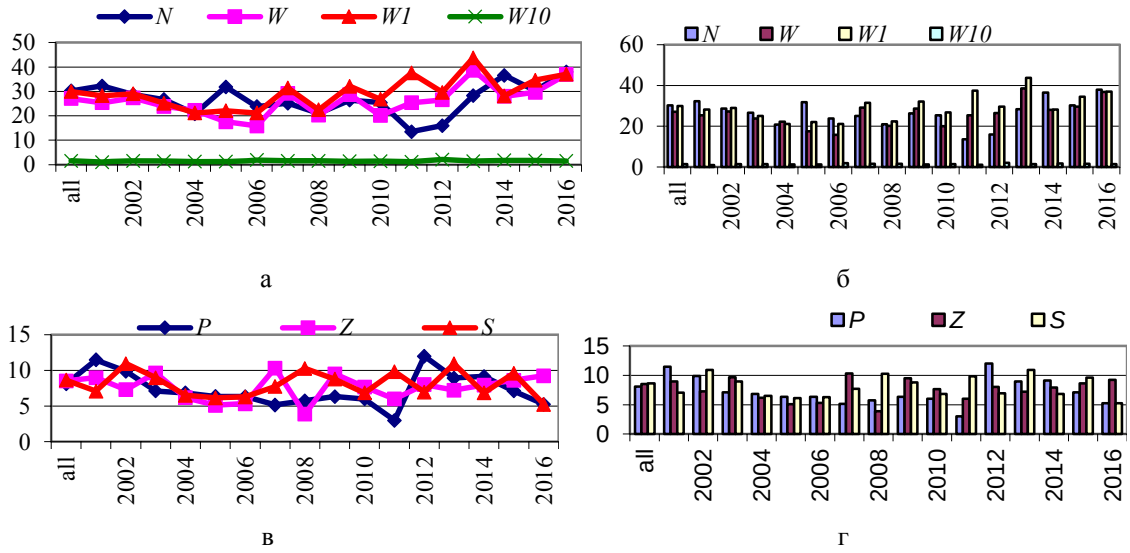


Рисунок 8 – Результат аналізу зміни динаміки вживання слів в досліджуваних текстах за визначений період: а – динаміка зміни параметрів мовлення в часі; б – розподіл значень параметрів мовлення за обумовлений період досліджень; в – динаміка зміни вживання сполучень, прийменників та речень в досліджуваних текстах; г – розподіл значень вживання сполучень, прийменників та речень за визначений період досліджень стилів автора

Довели, що існує динаміка зміни не лише коефіцієнтів мовлення авторського тексту за визначений період його творчості. Також є динаміка зміни і окремих складових, як кількість вживання словоформ на загальну кількість слів, сполучників та прийменників, речень у визначеному обсязі уривку, словоформ, які вживані лише один раз, та які вживані понад 10 разів.

6. ОБГОВОРЕННЯ

Для більш точного визначення величини приросту кожного із досліджуваного параметру необхідно провести більш суттєве дослідження на більшій вибірці як самих одноосібних творів, так збільшити діапазон дослідження творчості різних авторі на більший часовий проміжок творчості.

Далі проаналізуємо вибірку за авторським стилем та оберемо найкращий алгоритм для визначення стилю автора. На рис. 9а графік відображає визначення стилю автора по коефіцієнтах мовлення. На рис. 9б графік із накопиченням відображає зміни загальної суми за коефіцієнтами мовлення. На рис. 9в нормований графік відображає зміну вкладення кожного значення за коефіцієнтами мовлення.

Як бачимо, коефіцієнти авторського мовлення окрім *Kz* значно не змінюються в залежності від стилю конктерного автора для україномовних науково-технічних текстів. Або зміни є в млих межах, що ускладнює процес ідентифікації особливостей стилю мовлення конкретного автора в множині аналізованих авторських стилів. І чим більшою є така множина, тим складнішою буде процес ідентифікації

DOI 10.15588/1607-3274-2020-1-12

стилю конкретного автора без додаткових параметрів аналізу. Тоді проаналізуємо вибірку за авторським стилем за додатковими параметрами як загальна кількість речень в однакох за обсягом уривків, кількість слів у вибірці, частотність та пояка прийменників та сполучників. На рис. 10 графік відображає визначення стилю автора по додаткових параметрах авторського мовлення.

На рис. 10б графік із накопиченням відображає зміни загальної суми за параметрами. На рис. 10в нормований графік відображає зміну вкладення кожного значення за параметрами. Як бачимо введення додаткови параметрів зменшить множину авторів, стилі мовлення яких подібні для україномовного науково-технічного стилю публікацій. Введмо ще додаткові параметри як кількість речень, сполучників та прийменників (рис. 11) та проаналізуємо динаміку (табл. 4).

В табл. 4 наведені результати аналізу стилю 94 авторів на одноосібних працях (понад 200 одноосібних робіт) технічного спрямування за період 2001–2017 рр. Для кожного автора виведено середньоарифметичне значення кожного коефіцієнта та параметра мовлення на основі аналізу декілької його робіт за цей визначений період. Також проаналізовані стилі 4-х статтей одного авторського колективу під № 1–4 (в таблиці виділено жирним), частина авторів яких є в табл. 4 під № 69 та 93 (в таблиці виділено курсивом). Однак замала вибірка текстів для аналізу (понад 200) та кількості авторів (94) не гарантує точних результатів. Дослідження має бути продовжене на більшій кількості текстів, до яких незавжди маємо доступ. В подальшому необхідно також вдосконалити метод зарахунок аналізу текстів методами стилем атрії та глотохронології.

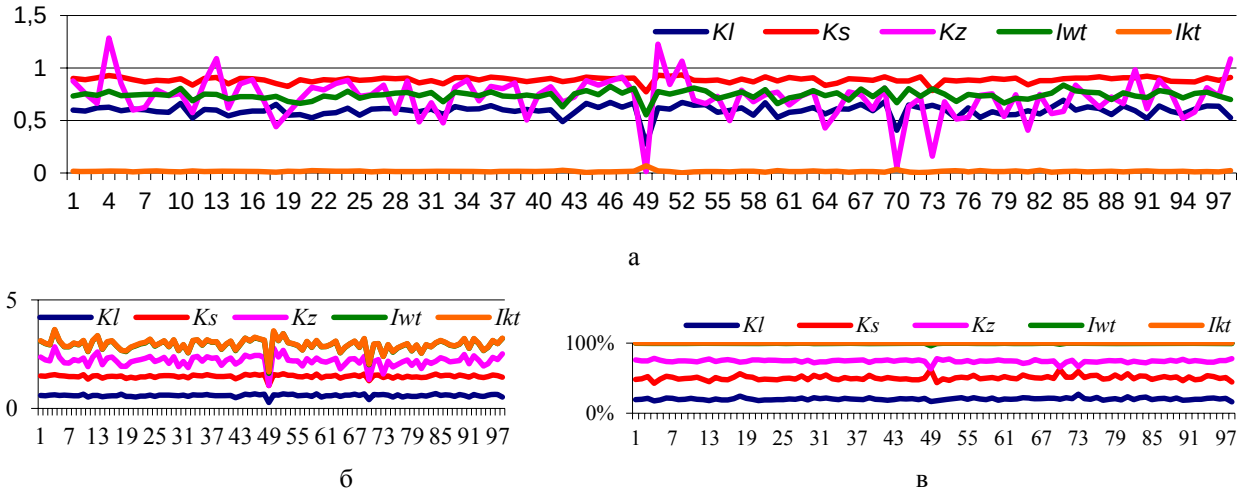


Рисунок 9 – Детальний аналіз: а – процесу у часі визначення стилю автора по коефіцієнтах мовлення; б – зміни загальної суми за коефіцієнтами мовлення; в – зміни вкладення кожного значення за коефіцієнтами мовлення

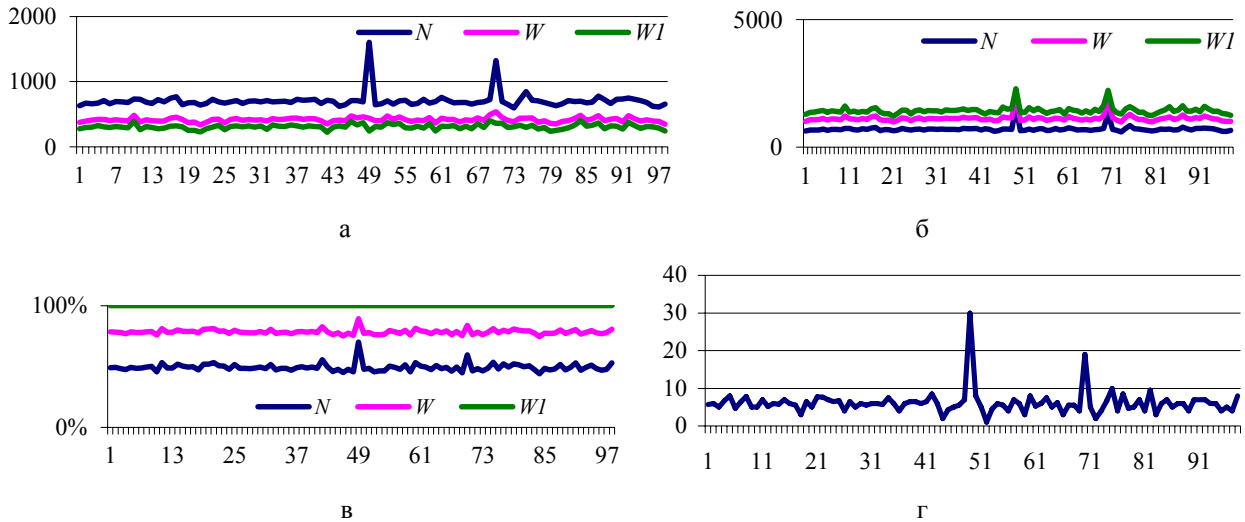


Рисунок 10 – Детальний аналіз: а – процесу визначення стилю автора по параметрах мовлення; б – зміни загальної суми за коефіцієнтами мовлення; в – зміни вкладення кожного значення за коефіцієнтами мовлення; г – зміни параметру як частота появи слова понад 10 разів (W_{10})

DOI 10.15588/1607-3274-2020-1-12

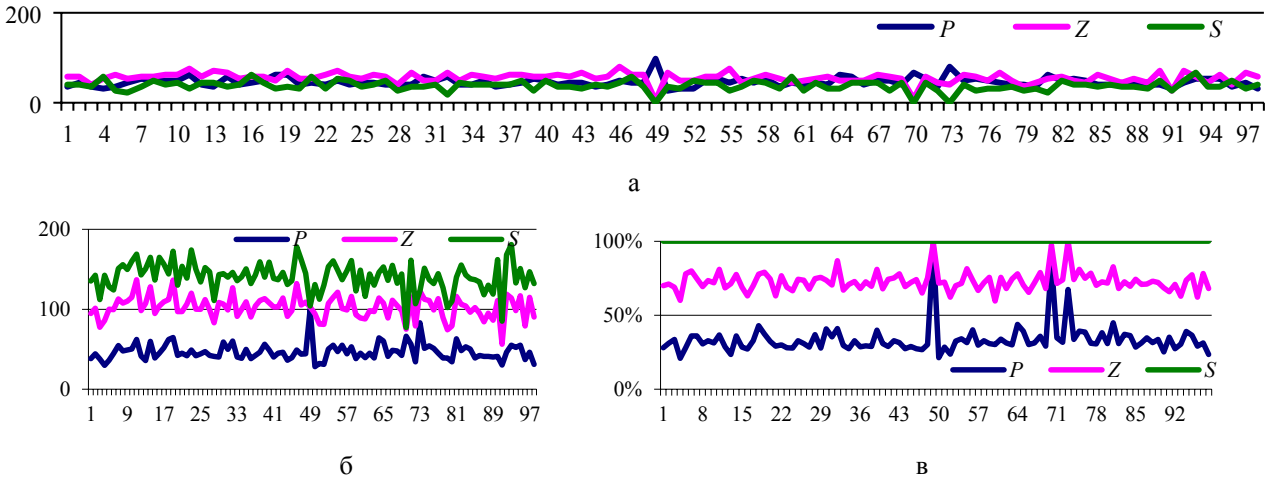


Рисунок 11 – Детальний аналіз: а – процесу визначення стилю автора по параметрах мовлення; б – зміни загальної суми за коефіцієнтами мовлення; в – зміни вкладення кожного значення за коефіцієнтами мовлення

Таблиця 4 – Результат роботи алгоритму аналізу стилю автора публікації на інформаційному ресурсі Victana

№	N	S	P	Z	W	W_1	W_{10}	I_{kt}	I_{wt}	K_z	K_s	K_l
1.	**631,3**	**40,7**	**38**	**56,7**	**377,7**	**277,7**	**5,7**	**0,015**	**0,73**	**0,88**	**0,9**	**0,6**
2.	**658**	**42**	**31**	**59**	**345**	**241**	**8**	**0,023**	**0,7**	**1,07**	**0,91**	**0,52**
3.	**614**	**32**	**46**	**69**	**391**	**287**	**4**	**0,01**	**0,73**	**0,73**	**0,88**	**0,64**
4.	**622**	**48**	**37**	**42**	**397**	**305**	**5**	**0,013**	**0,77**	**0,81**	**0,91**	**0,64**
5.	680	34	55	62	414	314	4	0,01	0,76	0,58	0,87	0,6
6.	709	35	52	46	398	285	6	0,015	0,72	0,52	0,87	0,56
7.	732	67	55	59	429	329	6	0,014	0,77	0,76	0,87	0,59
8.	749	49	46	73	478	375	7	0,015	0,78	0,88	0,9	0,64
9.	734	29	30	26	381	273	7	0,018	0,72	0,61	0,92	0,52
10.	730	51	41	70	433	317	7	0,016	0,73	0,98	0,91	0,59
11.	665	33	40	46	425	324	4	0,009	0,76	0,66	0,91	0,64
12.	723	35	41	54	401	280	6	0,015	0,7	0,72	0,9	0,55
13.	780	34	41	43	479	366	6	0,013	0,76	0,63	0,91	0,61
14.	685	39	42	53	432	333	5	0,012	0,77	0,73	0,9	0,63
15.	674	35	39	63	404	316	7	0,017	0,78	0,84	0,9	0,9
16.	700	42	50	46	485	406	6	0,012	0,84	0,59	0,9	0,69
17.	695	39	53	51	436	332	3	0,007	0,76	0,57	0,88	0,63
18.	709,5	49,5	48	58	399	292,5	9,5	0,024	0,73	0,75	0,88	0,56
19.	661	24	63	53	391	275	4	0,01	0,7	0,41	0,84	0,59
20.	631	31	34	45	350	249	7	0,02	0,71	0,75	0,9	0,55
21.	654	28	39	35	361	240	5	0,014	0,66	0,54	0,89	0,55
22.	682,3	37,7	39	50,3	398,7	296,3	4,7	0,012	0,74	0,75	0,9	0,58
23.	706	31	45	68,5	374	275	8,5	0,023	0,73	0,74	0,88	0,53
24.	712,5	33	51	48	442,5	331,5	4	0,009	0,75	0,53	0,88	0,62
25.	846	26	54	57	440	299	10	0,023	0,68	0,51	0,88	0,52
26.	726,3	39	51	61,3	441,3	332,3	6,7	0,015	0,75	0,68	0,88	0,6
27.	598	0	83	40	386	309	4	0,01	0,8	0,16	0,78	0,65
28.	652	28	34	45	405	296	2	0,005	0,73	0,72	0,92	0,62
29.	697	46	56	59,5	450	361,5	5	0,011	0,8	0,63	0,88	0,65
30.	1325	2	66	9	538	360	19	0,035	0,67	0,06	0,88	0,4
31.	726	46	42	56	493	399	4	0,008	0,81	0,81	0,91	0,68
32.	689,5	28	47,5	57	407,5	296	5,5	0,014	0,73	0,61	0,88	0,59
33.	683	43,5	48,5	63	446	357	5,5	0,012	0,8	0,74	0,89	0,65
34.	658	47	41	48	399	277	3	0,008	0,69	0,78	0,9	0,6
35.	682,6	45	60	47,8	416,2	318	6,2	0,015	0,76	0,59	0,86	0,6
36.	679	32	64	50	381	280	5	0,013	0,73	0,43	0,83	0,56
37.	673,5	33	39	58	419	329	7,5	0,018	0,79	0,78	0,91	0,62
38.	717	46	45	53	422	310	6	0,014	0,73	0,73	0,89	0,59
39.	761	28,3	39,3	48,5	440	315,8	5,3	0,012	0,71	0,65	0,91	0,58
40.	693	60	45	44	366	242	8	0,022	0,66	0,77	0,88	0,53
41.	670	30	38	55	449	356	3	0,007	0,79	0,75	0,92	0,67
42.	732	45	53	63	402	290	6	0,015	0,72	0,68	0,87	0,55
43.	666	49	44	55	412	318	7	0,017	0,77	0,79	0,89	0,62
44.	652	36	55	46	389	287	4	0,01	0,74	0,5	0,86	0,6
45.	716	27,5	47	74,5	413,5	293	5,5	0,013	0,71	0,73	0,89	0,58
46.	704,8	45,8	54,8	60	458,8	360	6	0,013	0,78	0,66	0,88	0,65

DOI 10.15588/1607-3274-2020-1-12

№	*N*	*S*	*P*	*Z*	*W*	W_1	W_{10}	I_{kt}	I_{wt}	K_z	K_s	K_l
47.	656	46	50	57,5	422,5	341,5	4,5	0,011	0,81	0,69	0,88	0,64
48.	705	49	31	50	474	369	1	0,002	0,78	1,06	0,93	0,67
49.	661,5	31	32	49,5	402,5	302	5	0,012	0,75	0,84	0,92	0,6
50.	644	37	28	66	400	310	8	0,02	0,78	1,23	0,93	0,62
51.	1602	1	100	3	442	245	30	0,068	0,55	0,01	0,77	0,28
52.	689	36	44	65	458	369	7	0,015	0,81	0,77	0,9	0,66
53.	708	56,5	43,5	62	442,5	336,5	5,5	0,012	0,76	0,91	0,9	0,63
54.	708	46	49	83	475	392	5	0,011	0,83	0,88	0,9	0,67
55.	645	37,7	39,3	58,7	403	302,3	4,3	0,011	0,74	0,84	0,9	0,62
56.	620	40	36	55	411	323	2	0,005	0,79	0,88	0,91	0,66
57.	699	32	46	68	401	302	6	0,015	0,75	0,72	0,89	0,57
58.	715,5	34	45	58	352	223,5	8,5	0,024	0,63	0,68	0,87	0,49
59.	666	35,5	40	63	401,5	305	6,5	0,016	0,76	0,82	0,9	0,6
60.	728	51	49	59	430	313	6	0,014	0,73	0,75	0,89	0,59
61.	717,5	26,5	56	57,5	433,5	321,5	6,5	0,015	0,74	0,5	0,87	0,6
62.	714,5	48,5	46	65	418,5	304,5	6,5	0,016	0,73	0,86	0,89	0,59
63.	730	39	42	62	440	323	6	0,014	0,73	0,8	0,9	0,6
64.	683	42	38	52	438	339	4	0,009	0,77	0,82	0,91	0,64
65.	699	41	49,5	60	427	314	6	0,014	0,74	0,69	0,88	0,61
66.	695	41	38,5	61,3	422,5	318,3	7,5	0,018	0,75	0,89	0,91	0,6
67.	691	44,7	40	51	436,7	336,7	5,7	0,013	0,77	0,82	0,91	0,63
68.	711	19	60	67	396	268	6	0,015	0,68	0,48	0,85	0,56
69.	*688,8*	*41,3*	*49,7*	*49,3*	*416,8*	*321,9*	*6*	*0,016*	*0,77*	*0,67*	*0,88*	*0,6*
70.	704,5	38	59	47,5	412	303,5	5,5	0,013	0,74	0,49	0,86	0,58
71.	700	35	40	68,5	418,5	320,5	6	0,014	0,77	0,88	0,9	0,6
72.	665	28	41	42	406	309	5	0,012	0,76	0,57	0,9	0,61
73.	708,5	47,5	42	57,5	434	323,5	6,5	0,015	0,75	0,84	0,9	0,61
74.	691	40	47	65	421	311	4	0,01	0,74	0,74	0,89	0,6
75.	668,8	34,5	44	55,8	368,3	262,5	6,8	0,018	0,71	0,73	0,88	0,55
76.	691,7	50	41,8	58,2	425,7	331,3	6,5	0,015	0,78	0,88	0,9	0,62
77.	731	54	49	71	420	301	7	0,017	0,72	0,85	0,88	0,57
78.	665	32,3	41,7	65	376	275,7	7,7	0,02	0,73	0,79	0,89	0,57
79.	642	56,8	44,8	52,3	337,5	230,3	7,8	0,023	0,68	0,81	0,87	0,52
80.	680	33	42	55	379	251	5	0,013	0,66	0,7	0,89	0,56
81.	677,5	36	64,5	72	373,5	255	6,5	0,018	0,68	0,57	0,86	0,55
82.	647	32	62	50	422	308	3	0,007	0,73	0,44	0,85	0,65
83.	768	47	51,5	58	452,5	323	5,5	0,012	0,71	0,68	0,89	0,59
84.	745	61	45	59	439	319	6	0,014	0,73	0,89	0,9	0,59
85.	691	42,3	39	55,3	396,7	289	7	0,018	0,73	0,85	0,9	0,57
86.	724,2	36,8	59,6	68,4	394,2	278,8	5,8	0,015	0,71	0,61	0,85	0,55
87.	665,5	43	35,5	72	399	299	6	0,015	0,75	1,09	0,91	0,6
88.	686,5	45	41,1	56,9	414,5	312,6	5,9	0,012	0,75	0,86	0,9	0,6
89.	729	32	62	75	380	261	7	0,018	0,69	0,58	0,84	0,52
90.	733,5	45	50	65	486,5	392	5	0,01	0,8	0,76	0,9	0,66
91.	682,5	39,7	49	61	394,2	291	5	0,013	0,74	0,74	0,88	0,58
92.	691,8	47,8	47,8	60	403,4	301,6	7,8	0,019	0,75	0,79	0,88	0,58
93.	*694,5*	*38,1*	*54,3*	*58,5*	*417,4*	*313,1*	*6,4*	*0,015*	*0,75*	*0,62*	*0,87*	*0,6*
94.	661,1	24,8	44,7	54,7	402,7	299,7	4,7	0,012	0,74	0,6	0,89	0,61
95.	708	28	36	64	419	309	8	0,019	0,74	0,85	0,91	0,59
96.	668,8	57	29,8	56	418,3	325,8	6,8	0,016	0,78	1,28	0,93	0,63
97.	662,5	34,8	37,8	39,8	410,3	303	5	0,012	0,74	0,67	0,9	0,61
98.	671,3	41,1	44,2	57,1	395,6	299	6	0,015	0,76	0,76	0,89	0,59

ВИСНОВКИ

Розроблено метод визначення автора тексту на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту. Розроблено алгоритм лексичного аналізу україномовних текстів та алгоритм синтаксичного аналізатора текстового контенту на основі аналізу кожного слова з врахуванням його частини мови та відмінювання. Тобто при аналізі лінгвістичних одиниць типу слів, враховувалась належність до частини мови та відмінювання в межах цієї частни мови. Для цього прводився аналіз флекцій цих слів для класифікації, виділення основи для формування відповідних алфавітно-частотних словників. Наповнення цих словників в подальшому враховувалися на наступних кроках визначення авторства тексту як розрахунок параметрів та коефіцієнтів авторського мовлення. Для індивідуального стилю письменника показовими є саме службові (стопові або опорні) слова, оскільки вони ніяк не пов'язані з темою і змістом публікації. Розроблено алгоритм визначення стопових слів текстового контенту на основі лінгвістичного аналізу текстового контенту. Його особливостями є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текс-

DOI 10.15588/1607-3274-2020-1-12

тів. Наведено теоретичне та експериментальне обгрунтування методу контент-моніторінгу та визначення стопових слів україномовного тексту. Метод спрямовано на автоматичне виявлення значущих стопових слів україномовного тексту за рахунок запропонованого формального підходу до реалізації парсингу текстового контенту науково-технічного спрямування. Запропоновано підхід до розроблення програмного забезпечення контент-моніторінгу для визначення автора в україномовних науко-технічних текстах на основі NLP, стилементії та Web Mining. Проаналізовано розробленою системою понад 200 одноосібних наукових публікацій зі всіх номерів Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі» (Україна) за період 2001–2017 рр. Досліджено внутрішню «динаміку» цих текстів довільно обраних авторів через аналіз коефіцієнтів зв'язності мовлення, лексичної різноманітності та синтаксичної складності, а також індексів концентрації та винятковості для перших *k*, *n* та *m* (без заголовка) слів авторського уривку та аналізованого.

Досліджено результати експериментальної апробації запропонованого методу контент-моніторінгу для визначення автора в україномовних наукових текстах технічного профілю. Проведено порівняння результатів на множині 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу. На основі розробленого програмного забезпечення отримано результати експериментальної апробації запропонованого методу контент-моніторінгу для визначення та аналізу стопових слів в україномовних наукових текстах технічного профілю на основі технології Web Mining. Виявлено, що для обраної експериментальної бази з понад 200 робіт найкращих результатів за критерієм щільності досягає метод аналізу статті без початкової обов'язкової інформації як анотації та ключові слова різними мовами, а також списку літератури. Подальшого експериментального дослідження потребує апробація запропонованого методу для визначення стилю автора з інших категорій текстів – наукових гуманітарного профілю, художніх, публіцистичних тощо.

ПОДЯКИ

Роботу виконано в рамках держбюджетної теми «Методи та засоби функціонування систем підтримки прийняття рішень на основі онтологій» (ID:839 2017-05-15 09:20:01 (2459-315)). Дослідження провадилось в межах спільних наукових досліджень кафедри інформаційних систем та мереж НУ «Львівська політехніка» на тему «Дослідження, розроблення і впровадження інтелектуальних розподілених інформаційних технологій та систем на основі ресурсів баз даних, сховищ даних, просторів даних та знань з метою прискорення процесів формування сучасного інформаційного суспільства». Наукові дослідження провадилися також в рамках ініціативної тематики досліджень кафедри ІСМ НУ «Львівська політехніка» на тему «Розроблення інтелектуальних розподілених систем на основі онтологічного підходу з метою інтеграції інформаційних ресурсів».

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Mobasher B. Data mining for web personalization / B. Mobasher // The adaptive web. – 2007. – Vol. 4321. – P. 90–135.
2. Dinucă C. Web Content Mining. In: University of Petroşani / C. Dinucă, D. Ciobanu // Economics. – 2012. – Vol. 12. – P. 85–92.
3. Xu G. Web content mining / G. Xu, Y. Zhang, L. Li // Web Mining and Social Networking. – 2011. – Vol. 6. – P. 71–87.
4. Khribi M. K. Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval / M. K. Khribi, M. Jemni, O. Nasraoui // Advanced Learning Technologies : International Conference, 1–5 July 2008 : proceedings. – Santander, Cantabria, Spain : IEEE, 2008. – P. 241–245.
5. Automatic web content personalization through reinforcement learning / [S. Ferretti, S. Mirri, C. Prandi, P. Salomoni] // Journal of Systems and Software. – 2016. – Vol. 121. – P. 157–169.
6. User attitudes towards news content personalization / [T. Lavie, M. Sela, I. Oppenheim et al] // International journal of human-computer studies. – 2010. – Vol. 68(8). – P. 483–495.
7. Fredrikson M. Repriv: Re-imagining content personalization and in-browser privacy / M. Fredrikson, B. Livshits // Symposium on Security and Privacy: Conference, 22–25 May 2011 : proceedings. – Berkeley, CA, USA : IEEE, 20011. – P. 131–146.
8. Application of neural networks and Kano's method to content recommendation in web personalization / [C. C Chang, P. L. Chen, F. R. Chiu, Y. K. Chen] // Expert Systems with Applications. – 2009. – Vol. 36(3). – P. 5310–5316.
9. Pat. US7,571,226B1 US Content personalization over an interface with adaptive voice character / [H. Partovi, R. Brathwaite, A. Davis et al.] (US) ; TellMe Networks, Inc., Mountain View, CA (US). – No.: 09/523,853 ; Marz 14, 2009; August 4, 2009, Patent and Trademark Office. – 20 p.
10. Pat. US2009/0171968A1 US Widget-assisted content personalization based on user behaviors tracked across multiple web sites / F. J. Kane, C. Hicks (US) ; Amazon Technologies Inc (US). – No.: 11/966,817 ; December 28, 2007; Jule 2, 2009, Google Patents. – 24 p.
11. Mirri S. Experiential adaptation to provide user-centered web content personalization / S. Mirri, C. Prandi, P. Salomoni // Advances in Human oriented and Personalized Mechanisms, Technologies, and Services : The Sixth International Conference, October 27 – November 1, 2013: proceedings. – Venice, Italy : IARIA, 2003. – P. 31–36.
12. Fernandez-Luque L. Review of extracting information from the Social Web for health personalization / L. Fernandez-Luque, R. Karlsen, J. Bonander // Journal of medical Internet research. – 2011. – Vol. 13(1). – P. 15.

DOI 10.15588/1607-3274-2020-1-12

13. Pat. US8,019,777B2 US Digital content personalization method and system / E. Hauser (US) ; CRICKET MEDIA Inc (US). –No.: 12/795,419 ; June 7, 2010; September 13, 2011, Patent and Trademark Office. – 15 p.
14. Ho S. Y. Timing of adaptive web personalization and its effects on online consumer behavior / S. Y. Ho, D. Bodoff, K. Y. Tam // Information Systems Research. – 2011. – Vol. 22(3). – P. 660–679.
15. Uchyigit G. Personalization techniques and recommender systems / G. Uchyigit, M. Y. Ma. – Singapore : World Scientific, 2008. – 322 p.
16. Pat. US2006/0020883A1 Web page personalization / [N. Kothari, M. Harder, R. Howard et al.] (US) ; Microsoft Technology Licensing LLC (US). – No.: 10/857,724 ; May 28, 2004; Januar 26, 2006, Patent and Trademark Office. – 18 p.
17. Zhang H. Construction of ontology-based user model for web personalization / H. Zhang, Y. Song., H. T. Song // Lecture Notes in Computer Science. – 2007. – Vol. 4511. – P. 67–76.
18. Pat. US 8,254,892 B2 US Methods and apparatus for anonymous user identification and content personalization in wireless communication / H. Chien (US) ; AT&T Mobility II LLC (US). – No.: 12/468,708 ; September 10, 2009; August 28, 2012, Patent and Trademark Office. – 9 p.
19. Pat. US7,970,664B2 US Content personalization based on actions performed during browsing sessions / G. D. Linden, B. R. Smith, N. K. Zada (US) ; Amazon Technologies Inc (US). – No.: 11/009,732 ; December 10, 2004; June 28, 2011, Patent and Trademark Office. –36 p.
20. Web personalization using web mining: concept and research issue / [P. Mehtaa, B. Parekh, K. Modi, P. Solanki] // International Journal of Information and Education Technology. – 2012. – Vol. 2(5). – P. 510–512.
21. Zhezhnych P. Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects / P. Zhezhnych, O. Markiv // Advances in Intelligent Systems and Computing. – 2018. – Vol. 689. – P. 656–667.
22. Basyuk T. The main reasons of attendance falling of internet resource / T. Basyuk // Computer Sciences and Information Technologies : Xth International Scientific and Technical Conference, 14–17 September 2015 : proceedings. – Lviv : IEEE, 2015. – P. 91–93.
23. Uniform Method of Operative Content Management in Web Systems / [A. Gozhyj, L. Chyrun, A. Kowalska-Styczen, O. Lozynska] // CEUR Workshop Proceedings. – 2018. – Vol. 2136. – P. 62–77.
24. Kravets P. The control agent with fuzzy logic / P. Kravets // Perspective Technologies and Methods in MEMS Design : VIth International Conference, 20–23 April 2010 2015 : proceedings. – Lviv : IEEE, 2015. – P. 40–41.
25. Davydov M. Linguistic Models of Assistive Computer Technologies for Cognition and Communication / M. Davydov, O Lozynska // Computer Science and Information Technologies : XIth International Scientific and Technical Conference, 6–10 September 2016 : proceedings. – Lviv : IEEE, 2016. – P. 171–175.
26. Mykich K. Algebraic model for knowledge representation in situational awareness systems / K. Mykich, Y. Burov // Computer Sciences and Information Technologies : nternational Scientific and Technical Conference, 6–10 September 2016 : proceedings. – Lviv : IEEE, 2016. – P. 165–167.
27. Mykich K. Uncertainty in situational awareness systems / K. Mykich, Y. Burov // Modern Problems of Radio Engineering, Telecommunications and Computer Science : 13th International Conference, 623–26 Februar 2016 : proceedings. – Lviv : IEEE, 2016. – P. 729–732.
28. Mykich K. Algebraic Framework for Knowledge Processing in Systems with Situational Awareness / K. Mykich, Y. Burov // Advances in Intelligent Systems and Computing. – 2017. – Vol. 512. – P. 217–227.
29. Mykich K. Research of uncertainties in situational awareness systems and methods of their processing / K. Mykich, Y. Burov // EasternEuropean Journal of Enterprise Technologies. – 2016. – Vol. 1(79). – P. 19–26.
30. Vysotska V. Linguistic Analysis of Textual Commercial Content for Information Resources Processing / V. Vysotska // Modern Problems of Radio Engineering, Telecommunications and Computer Science : 13th International Scientific and Technical Conference, 23–26 February 2016 : proceedings. – Lviv : IEEE, 2016. – P. 709–713.
31. Information resources processing using linguistic analysis of textual content / [J. Su, V. Vysotska, A. Sachenko, V. Lytvyn, Y. Burov] // Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications : 9th International Conference, 21–23 September 2017 : proceedings. – Bucharest, Romania: IEEE, 2017. – P. 573–578.
32. Content Linguistic Analysis Methods for Textual Documents Classification / [V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, H Rishnyak] // Computer Science and Information Technologies : 11th International Scientific and Technical Conference, 6–10 September 2016 : proceedings. – Lviv : IEEE, 2016. – P. 190–192.
33. Bisikalo O. V. Identifying keywords on the basis of content monitoring method in ukrainian texts / O. V. Bisikalo, V. A. Vysotska // Radio Electronics, Computer Science, Control. – 2016. – Vol. 1(36). – P. 74–83.
34. Bisikalo O.V. Sentence syntactic analysis application to keywords identification Ukrainian texts / O. V. Bisikalo, V. A. Vysotska // Radio Electronics, Computer Science, Control. – 2016. – Vol. 3(38). – P. 54–65.
35. Lytvyn V. Application of algorithmic algebra system for grammatical analysis of symbolic computation expressions of propositional logic / V. Lytvyn, I. Bobyk, V. Vysotska // Radio Electronics, Computer Science, Control. – 2016. – Vol. 4(39).– P. 54–67.
36. Alieksieieva K. Technology of commercial web-resource management based on fuzzy logic / K. Alieksieieva, A. Berko, V. Vysotska // Radio Electronics, Computer Science, Control. – 2015. – Vol. 3(34). – P. 71–79.
37. Application of Sentence Parsing for Determining Keywords In Ukrainian Texts / [Vasyl Lytvyn, Victoria Vysotska, Dmytro Dosyn, Roman Holoschuk, Zoriana Rybchak] // Computer Science and Information Technologies : 12th International Scientific and Technical Conference, 5–8 September 2017 : proceedings. – Lviv : IEEE, 2017. – P. 326–331.

Received 25.10.2019.
Accepted 09.02.2020.

DOI 10.15588/1607-3274-2020-1-12

УДК 004.9

МЕТОД АВТОРИФИКАЦИИ ТЕКСТА НАУЧНО-ТЕХНИЧЕСКИХ ПУБЛИКАЦИЙ НА ОСНОВЕ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА КОЭФФИЦИЕНТОВ ЯЗЫКОВОГО РАЗНООБРАЗИЯ

Высоцкая В. А. – канд. техн. наук, доцент, доцент кафедры «Информационные системы и сети», Национальный университет «Львовская политехника», Украина.

АННОТАЦИЯ

Актуальность. Авторификация авторства текста является техникой определения автора текста, когда неоднозначно, кто ее написал. Это полезно, когда несколько человек претендуют на авторство одной публикации или в случаях, когда никто не претендует на авторство текстового контента, например, так называемые тролли в социальных сетях во время информационной войны. Сложность проблемы авторского текста, очевидно, экспоненциально выше, большее количество возможных авторов. Наличие авторских текстовых образцов также является существенным при продвижении этой проблемы. Атрибуция авторского текста включает следующие три проблемы:

– выявление автора текстового автора из группы возможных или ожидаемых авторов, где автор всегда находится в группе подозреваемых;

– не идентификация автора текстового автора из группы возможных или ожидаемых авторов, где автор может не быть в группе подозреваемых;

– оценка возможности данного текста, написанного данным автором или нет.

Поэтому задача автоматического определения автора текстового контента научно-технического направления актуальна и требует новых (более совершенных) подходов к ее решению.

Целью исследования является разработка метода определения автора в украиноязычных текстах на основе технологии лингвометрии.

Метод. Разработано лингвометричний метод алгоритмического обеспечения процессов контент-мониторинга для решения задачи автоматического определения автора русскоязычного текстового контента на основе технологии статистического анализа коэффициентов языкового разнообразия. Проведения декомпозиции метода определения автора на основе анализа таких коэффициентов речи как лексическая разнообразие, степень (мера) синтаксической сложности, связность речи, индексы исключительности и концентрации текста. Проанализированы также параметры авторского стиля как количество слов в определенном тексте, общее количество слов этого текста, количество предложений, количество предлогов, количество союзов, количество слов с частотой 1, количество слов с частотой 10 и больше. Особенностями разработанного является адаптация морфологического и синтаксического анализа лексических единиц к особенностям конструкций украиноязычных слов / текстов. То есть при анализе лингвистических единиц типа слов, учитывалась принадлежность к части речи и склонение в пределах этой части речи. Для этого проводился анализ флексий этих слов для классификации, выделение основы для формирования соответствующих алфавитно-частотных словарей. Наполнение этих словарей в дальнейшем учитывались на следующих шагах определения авторства текста как расчет параметров и коэффициентов авторской речи. Для индивидуального стиля писателя показательны именно служебные (стоп или опорные) слова, поскольку они никак не связаны с темой и содержанием публикации.

Результаты. Проведено сравнение результатов на множестве 200 самостоятельных работ технического направления около 100 различных авторов период 2001–2017 гг. Для определения меняются и как коэффициенты разнообразия текста этих авторов в разные промежутки времени.

Выводы. Выявлено, что для выбранной экспериментальной базы из более 200 работ лучших результатов по критерию плотности достигает метод анализа статьи без начальной обязательной информации как аннотации и ключевые слова на разных языках, а также список литературы.

КЛЮЧЕВЫЕ СЛОВА: текстовый контент, NLP, контент-мониторинг, стоп-слова, контент-анализ, статистический лингвистический анализ, квантитативных лингвистика.

УДК 004.9

THE SCIENTIFIC AND TECHNICAL PUBLICATIONS TEXT AUTHORITICATION METHOD BASED ON LINGUSTICAL ANALYSIS OF LANGUAGE DIVERSITY COEFFICIENTS

Vysotska V. – PhD, Associate Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. Authorization of the authorship of the text is a technique for determining the author of the text, when it is ambiguous who wrote it. It is useful when several people claim to be the authors of one publication or in cases where nobody claims to authorship of text content, for example, so-called trolls in social networks during an information warfare. The complexity of the problem of the author's text, obviously, is exponentially higher, more likely authors. The presence of author's text samples is also significant in advancing this problem. The attribution of the author's text includes the following three problems:

– author discovery of text from probable or expected authors group, where the author is always in a suspects group;

– not identification of the author of a text author from a group of probable or expected authors, where the author may not be in a group of suspects;

– assessment of the possibility of this text, written by the author or not.

Therefore, the task of automatically determining the author of text content of scientific and technical direction is relevant and requires new (more perfect) approaches to its solution.

Objective of the study is to develop a method for determining the author in Ukrainian texts based on the technology of lingometry.

DOI 10.15588/1607-3274-2020-1-12

Method. Lingvometric method of algorithmic provision of content monitoring processes for solving the problem of automatic determination of the author of Ukrainian-language text content on the basis of technology of statistical analysis of linguistic diversity coefficients is developed. A decomposition of the method of determination of the author on the basis of analysis of such broadcasting factors as lexical diversity, degree (degree) of syntactic complexity, speech connectivity, singularity indexes and text concentrations is made. Also, author's style parameters are analyzed as the number of words in a particular text, the total number of words in this text, the number of sentences, the number of prepositions, the number of conjunctions, the number of words with the frequency of 1, and the number of words with a frequency of 10 or more. The features of the developed is the adaptation of the morphological and syntactic analysis of lexical units to the features of the designs of Ukrainian-language words / texts. That is, in the analysis of linguistic units of the type of words, the affiliation with the part of speech and declarations within this part of the language was taken into account. To do this, an analysis of the flexion of these words was carried out for classification, the allocation of the basis for the formation of the corresponding alphabet-frequency dictionaries. The filling of these dictionaries was further taken into account in the subsequent steps of determining the authorship of the text as the calculation of parameters and coefficients of copyright broadcasting. For the individual style of a writer, it is precisely service (stop or reference) words that are indicative because they are not related to the topic and content of the publication.

Results. A comparison of results on a plurality of 200 individual technical works of about 100 different authors over the period 2001–2017 has been made to determine whether the coefficients of the diversity of the text of these authors are different at different intervals.

Conclusions. It has been found that for the chosen experimental base with over 200 works of the best results, the method of analysis of the article without initial obligatory information as annotations and keywords in various languages and the list of literature achieves the density criterion.

KEYWORDS: text content, NLP, content monitoring, stop words, content analysis, statistical linguistic analysis, quantitative linguistics.

REFERENCES

1. Mobasher B. Data mining for web personalization, *The adaptive web*, 2007, Vol. 4321, pp. 90–135.
2. Dinucă C., Ciobanu D. Web Content Mining. In: University of Petroşani, *Economics,* 2012, Vol. 12, pp. 85–92.
3. Xu G. Zhang Y., Li L. Web content mining, *Web Mining and Social Networking*, 2011, Vol. 6, pp. 71–87.
4. Khribi M. K., Jemni M., Nasraoui O. Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval, *Advanced Learning Technologies : International Conference, 1–5 July 2008 : proceedings.* Santander, Cantabria, Spain, IEEE, 2008, pp. 241–245.
5. Ferretti S., Mirri S., Prandi C., Salomoni P. Automatic web content personalization through reinforcement learning, *Journal of Systems and Software*, 2016, Vol. 121, pp. 157–169.
6. Lavie T., Sela M., Oppenheim I., Inbar O., Meyer J. User attitudes towards news content personalization, *International journal of human-computer studies,* 2010, Vol. 68(8), pp. 483–495.
7. Fredrikson M., Livshits B. Repriv: Re-imagining content personalization and in-browser privacy, *Symposium on Security and Privacy: Conference, 22–25 May 2011 : proceedings*. Berkeley, CA, USA, IEEE, 2011. pp. 131–146.
8. Chang. C., Chen P., Chiu F., Chen Y. Application of neural networks and Kano's method to content recommendation in web personalization, *Expert Systems with Applications,* 2009, Vol. 36(3), pp. 5310–5316.
9. Partovi H., Brathwaite R., Davis A., McCue M., Porter B., Giannandrea J., Li Z. (US) Pat. US7,571,226B1 US Content personalization over an interface with adaptive voice character, U.S. ; TellMe Networks, Inc., Mountain View, CA (US). No.: 09/523,853 ; Marz 14, 2009; August 4, 2009, Patent and Trademark Office, 20 p.
10. Kane F. J., Hicks C. (US) Pat. US2009/0171968A1 US Widget-assisted content personalization based on user behaviors tracked across multiple web sites; Amazon Technologies Inc (US). No.: 11/966,817; December 28, 2007; Jule 2, 2009, Google Patents, 24 p.
11. Mirri S., Prandi C., Salomoni P. Experiential adaptation to provide user-centered web content personalization, *Advances in Human oriented and Personalized Mechanisms, Technologies, and Services : The Sixth International Conference, October 27 – November 1, 2013: proceedings*. Venice, Italy, IARIA, 2003, pp. 31–36.
12. Fernandez-Luque L., Karlsen R., Bonander J. Review of extracting information from the Social Web for health personalization, *Journal of medical Internet research*, 2011, Vol. 13(1), P. 15.
13. Hauser E. (US) Pat. US8,019,777B2 US Digital content personalization method and system; CRICKET MEDIA Inc (US). No.: 12/795,419 ; June 7, 2010; September 13, 2011, Patent and Trademark Office, 15 p.
14. Ho S. Y., Bodoff D., Tam K. Y. Timing of adaptive web personalization and its effects on online consumer behavior, Information Systems Research, 2011, Vol. 22(3), pp. 660–679.
15. Uchyigit G., Ma M. Y.. Personalization techniques and recommender systems. Singapore, World Scientific, 2008, 322 p.
16. Kothari N., Harder M., Howard R., Sanabria A., Schackow S. (US) Pat. US2006/0020883A1 Web page personalization; Microsoft Technology Licensing LLC (US). No.: 10/857,724 ; May 28, 2004; Januar 26, 2006, Patent and Trademark Office. – 18 p.
17. Zhang H., Song Y., Song H. T. Construction of ontology-based user model for web personalization, *Lecture Notes in Computer Science,* 2007, Vol. 4511, pp. 67–76.
18. Chien H. (US) Pat. US 8,254,892 B2 US Methods and apparatus for anonymous user identification and content personalization in wireless communication; AT&T Mobility II LLC (US). No.: 12/468,708 ; September 10, 2009; August 28, 2012, Patent and Trademark Office. – 9 p.
19. Linden G. D., Smith B. R., Zada N. K. (US) Pat. US7,970,664B2 US Content personalization based on actions performed during browsing sessions; Amazon Technologies Inc (US). No.: 11/009,732 ; December 10, 2004; June 28, 2011, Patent and Trademark Office, 36 p.
20. Mehtaa P., Parekh B., Modi K., Solanki P. Web personalization using web mining: concept and research issue, *International Journal of Information and Education Technology*, 2012, Vol. 2(5), pp. 510–512.
21. Zhezhnych P., Markiv O. Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism

DOI 10.15588/1607-3274-2020-1-12

Documentation Objects, *Advances in Intelligent Systems and Computing*, 2018, Vol. 689, pp. 656–667.
22. Basyuk T. The main reasons of attendance falling of internet resource, *Computer Sciences and Information Technologies : Xth International Scientific and Technical Conference, 14–17 September 2015 : proceedings*. Lviv, IEEE, 2015, pp. 91–93.
23. Gozhyj A., Chyrun L., Kowalska-Styczen A., Lozynska O. Uniform Method of Operative Content Management in Web Systems, *CEUR Workshop Proceedings,* 2018, Vol. 2136, pp. 62–77.
24. Kravets P. The control agent with fuzzy logic, *Perspective Technologies and Methods in MEMS Design : VIth International Conference, 20–23 April 2010 2015 : proceedings.* Lviv, IEEE, 2015, pp. 40–41.
25. Davydov M., Lozynska O. Linguistic Models of Assistive Computer Technologies for Cognition and Communication, *Computer Science and Information Technologies : XIth International Scientific and Technical Conference, 6–10 September 2016 : proceedings.* Lviv, IEEE, 2016, pp. 171–175.
26. Mykich K., Burov Y. Algebraic model for knowledge representation in situational awareness systems, *Computer Sciences and Information Technologies : 11th International Scientific and Technical Conference, 6–10 September 2016 : proceedings.* Lviv, IEEE, 2016, pp. 165–167.
27. Mykich K., Burov Y. Uncertainty in situational awareness systems, *Modern Problems of Radio Engineering, Telecommunications and Computer Science : 13th International Conference, 623–26 Februar 2016 : proceedings.* Lviv, IEEE, 2016, pp. 729–732.
28. Mykich K., Burov Y. Algebraic Framework for Knowledge Processing in Systems with Situational Awareness, *Advances in Intelligent Systems and Computing*, 2017, Vol. 512, pp. 217–227.
29. Mykich K., Burov Y. Research of uncertainties in situational awareness systems and methods of their processing, *EasternEuropean Journal of Enterprise Technologies,* 2016, Vol. 1(79), pp. 19–26.
30. Vysotska V. Linguistic Analysis of Textual Commercial Content for Information Resources Processing, *Modern Problems of Radio Engineering, Telecommunications and Computer Science : International Scientific and Technical Conference, 23–26 February 2016 : proceedings*. Lviv, IEEE, 2016, pp. 709–713.
31. Su J., Vysotska V., Sachenko A., Lytvyn V., Burov Y. Information resources processing using linguistic analysis of textual content, *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications : 9th International Conference, 21–23 September 2017 : proceedings.* Bucharest, IEEE, 2017, pp. 573–578.
32. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. Content Linguistic Analysis Methods for Textual Documents Classification, *Computer Science and Information Technologies : 11th International Scientific and Technical Conference, 6–10 September 2016 : proceedings*. Lviv, IEEE, 2016, pp. 190–192.
33. Bisikalo O. V., Vysotska V. A. Identifying keywords on the basis of content monitoring method in ukrainian texts, *Radio Electronics, Computer Science, Control*, 2016, Vol. 1(36), pp. 74–83.
34. Bisikalo O. V., Vysotska V. A. Sentence syntactic analysis application to keywords identification Ukrainian texts, *Radio Electronics, Computer Science, Control,* Vol. 3(38), 2016, pp. 54–65.
35. Alieksieieva K., Berko A., Vysotska V. Technology of commercial web-resource management based on fuzzy logic *Radio Electronics, Computer Science, Control,* 2015, Vol. 3(34). pp. 71–79.
36. Lytvyn V., Bobyk I., Vysotska V. Application of algorithmic algebra system for grammatical analysis of symbolic computation expressions of propositional logic, *Radio Electronics, Computer Science, Control,* 2016, Vol. 4(39), pp. 54–67.
37. Lytvyn Vasyl, Vysotska Victoria, Dosyn Dmytro, Holoschuk Roman, Rybchak Zoriana Application of Sentence Parsing for Determining Keywords In Ukrainian Texts, *Computer Science and Information Technologies : 12th International Scientific and Technical Conference, 5–8 September 2017 : proceedings*. Lviv, IEEE, 2017, pp. 326–331.

DOI 10.15588/1607-3274-2020-1-12

UDC 004.627

DATA COMPRESSION IN BLACK-GRAY-WHITE BARCODING

Dychka I. – Dr. Sc., Professor, Head of the Faculty of Applied Mathematics, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.

Onai M. – PhD, Associate Professor in the Computer Systems Software Department, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.

Sulema Ol. – Post-graduate student of the Department of Computer Systems Software, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine.

ABSTRACT

Context. In this paper the authors propose a method for data compression to be used for presenting information in the form of 2D matrix barcode. The proposed method is based on both a structural-logical approach and using three colors in a barcode instead of two colors as it is in standard black-and-white barcodes. This approach allows to increase data density keeping the same area as bicolor barcodes take. In the paper authors present the data compression method and demonstrate the barcoding technology.

Objective. The goal of the work is to develop a method of data barcoding that would allow to encode more information in the form of 2D matrix barcode.

Method. The method of tricolor matrix barcoding with compression is proposed. The main idea of the method is to compress input textual information on the stage of alphanumeric sequence transformation into a set of barcode patterns, which will form a resulting barcode symbol. It is possible due to intermediate transformation of input characters from initial notation, which is determined by cardinality of an input alphabet, to a notation defined by cardinality of barcode patterns alphabet. Choice of the input alphabet influences overall compression, and it is an important step of the method to choose the initial alphabets for the textual information to be encoded. Use of three colors over standard two colors is also an important component for creating a barcode symbol with increased informational density. As ternary notation is used, the second transformation from the intermediate notation to the ternary one provides more compression. The proposed method allows to represent more textual data in a single barcode symbol than bicolor barcoding approaches do.

Results. The method of tricolor matrix barcoding with compression has been developed and described. Authors provided an example of the method implementation on test data that had been barcoded using the method.

Conclusions. The experiments conducted for this research have confirmed that the proposed method provides more informational density as compared with black-and-white matrix barcodes. The prospects for further research might include studying noise immunity issue in order to guarantee error-free scanning and increased reliability of the barcode, and extending the barcoding software to be used in any alphabet.

KEYWORDS: textual Data Protection, Automated Data Capture, Barcoding, 2D Barcode, Tricolor Barcode, Grayscale Barcode, Black-Gray-White Barcode.

ABBREVIATIONS

BGW-Code is a black-gray-white barcode;
DNM is a decimal numbers mode;
HNM is a hexadecimal numbers mode;
TDM is a textual data mode;
ASM is an ASCII symbols mode.

NOMENCLATURE

B is an area of barcode symbol;
s is a number of cells in a barcode pattern;
q is a number of colors, or number base;
V is a maximum capacity of a barcode symbol;
Ω is a symbolism (an alphabet) of barcode;
Ω_{inf} is an alphabet of informational (textual) symbols;
Ω_{aux} is an alphabet of auxiliary (technical) symbols;
P_X is a cardinality of any alphabet X;
T is an input alphanumeric sequence;
U is a resulting sequence of barcode patterns;
n is a number of adjacent symbols of the same type in the input sequence T;
m is a number of barcode patterns;
t_i is an element (one character) of the input sequence T;
w_i is a subsequence consisted of elements t_i;
ω_i is a barcode pattern;
u_z is a subsequence of barcode patterns;
$U^{(s)}_{P_{\Omega_{\text{inf}}}}(P_A)$ is a compression coefficient.

INTRODUCTION

One of the important problems in the field of modern information technologies is an issue of information support of automated relocation of objects (goods, freights, medical supplies, documents, parcels etc.) [1]. Automated objects relocation systems are based on using automated identification that enables entering required data into a computer system by its automated scanning from the object.

High demand for automated identification is determined by the desire to improve control on objects relocation, reduce production costs, and increase its profitability and efficiency.

Among various types of automated identification, it is the barcoding technology [2, 3] that has been widely distributed. It happens due to the low cost of both barcode patterns production and scanning equipment.

DOI 10.15588/1607-3274-2020-1-13

Barcoding is the way to represent and store information on a carrier using elementary discrete graphical shapes, such as circle, ellipse, square, rectangle, hatch (straight, oblique), triangle, polygon (hexagon, octagon) etc. Information is represented in the form of combinations of elements with different coloring.

Barcoding provides an optical way of information scanning, including distant scanning [4, 5]. A barcode is placed on an object surface, and it is moving along with the object throughout its trajectory.

Since the invention of barcoding (the first patent for the barcode was received in 1952), it has been more than 60 years, however barcoding is still considered as one of the advanced technologies. Moreover, a lot of experts believe that barcodes are among the most prominent discoveries of the 20th century.

There is a number of barcode types which can be divided into 3 main groups: linear, stack, and matrix. A few up to several dozens of alphanumeric codes can be represented in the form of linear barcode. Several hundred characters can be represented in the form of stack barcode. Up to several thousand symbols can be represented in the form of matrix barcode. The subject of this research is matrix barcoding.

Matrix barcode is a two-dimensional array of discrete graphic items combined as one image. The structure of such an array is called a barcode pattern (BC-pattern). The majority of matrix barcodes are black-and-white. However, in recent years there has been growing interest in the development of multicolor matrix barcodes. The most well-known among them are Microsoft's High Capacity Color Barcode [6] and High Capacity Color QR code [7].

Multicolor barcodes provide larger data density in comparison with black-and-white equivalents; however, they have a disadvantage, which is narrow scope of application. This is due to as yet higher cost of color printers compared to black-and-white printers, as well as the high cost of consumables for color printing. Therefore, multicolor barcodes are unable to replace black-and-white barcodes in all areas of applications, considering that black-and-white printing can be more efficient than a color one in certain use cases.

Since modern printing equipment, namely laser printers, provide a high black-and-white printing quality with the required resolution, we consider that it is appropriate to add one more color, the gray, representation of which is not a challenge using black-and-white printer. Thus, we propose to create black-gray-white barcode patterns (Fig. 1) with existing black-and-white printing equipment. As a result, it is expected that matrix barcode data density will be increased.

Subsequently, let us call such black-gray-white barcodes a BGW-Code.

However, to achieve high rate of data density, it is essential to apply a structural-logical approach for increasing barcodes density, in addition to the use of the third color.

Figure 1 – An example of BGW-Code symbol

The object of study is the process of transforming textual information into a barcode.

The process of a barcode symbol creation becomes more complex because of procedure of data compressing, which is an important part of the method proposed in the paper. Therefore, the process of data barcoding consists of two stages: compression of input information, and transforming compressed data into a barcode symbol.

The subject of study is the methods for data compression and barcode construction.

The methods being developed by authors are aimed at providing a possibility to encode more data into a single barcode symbol.

The purpose of the work is to develop the method of forming barcode symbol with increased informational density.

1 PROBLEM STATEMENT

Requirements to a barcode as a way to store and input information are as follows:

1. Miniaturization of barcode pattern (limited area B is allocated for a barcode pattern ω_i superimposing on an item).

2. Significantly increasing capacity V of a barcode pattern ω_i without changing its geometrical dimensions.

3. Widening the concept of a barcode pattern in order to obtain portable data file (the barcode pattern has to contain not only an access key to information but complete information about the item).

Let us consider a problem of increasing data density of a matrix barcode. Factors of increasing data density could be both a number of colors q used in barcode and the use of specific methods for data compression, which enable increasing capacity V of a barcode pattern ω_i.

In this research we consider tricolor barcode, i.e. $q = 3$, in which black, white, and gray colors are used for elements representation. Such tricolor barcode patterns are easily produced by using an ordinary printer.

DOI 10.15588/1607-3274-2020-1-13

Increasing data density basically means increasing of a ratio between an initial data sequence needed to be encoded and a resulting sequence of barcode patterns, which represents by a compression coefficient *U*.

Therefore, the formal definition of the task to be solved in this research is as follows:

$$\begin{cases} B(\omega_i) \to \min \\ V(\omega_i) \to \max, \forall i, \omega_i \in \Omega_{\text{inf}} \\ U^{(s)}_{P_{\Omega_{\text{inf}}}}(P_A) \to \max \end{cases}.$$

Thus, in this paper we present the new method of alphanumeric data compression; these data are a subject of representation in the form of black-gray-white barcode.

2 REVIEW OF THE LITERATURE

Methods of information barcoding as well as barcodes themselves are the subject of research for many scientists.

In [8], the author presents a method to generate and decode two-dimensional color barcode consisted of several blocks, which are a black-and-white configuration block that encodes auxiliary information about the barcode itself and a set of color data blocks that encodes actual data.

In the patent [9] it is proposed to store information decoded from a barcode in a form of character-based data in an auxiliary field (e.g. a comment field).

The authors of [10] propose a new approach of decoding color barcode, which does not require a reference color palette. They describe an algorithm, in which groups of color bars are decoded at once, what is exploiting the fact that joint color changes can be represented by a low-dimensional space.

A prototype for generating and reading a HCC2D code format on both PC and mobile phones is presented in [11]. The authors provide experimental results considering different operating scenarios and data densities in comparison with 2-dimensional barcodes.

The authors of [12] describe a method of high capacity color barcodes generation, which operates due to embedding independent data into two different printer colorant channels via halftone-dot orientation modulation.

In [13], an approach for localization and segmentation of a 2D color barcode when it is read using computer vision techniques is presented. The authors develop a progressive strategy to achieve high accuracy in diverse scenarios and computational efficiency.

The authors in [14] propose both a system and a method to encode and decode data in a color barcode pattern using dot orientation and color separability. They aver the method to be robust against interseparation misregistration with a small symbol error rate.

COBRA system, which is a visible light communication (VLC) system for off-the-shelf smartphones, is presented in [15]. The proposed system is able to encode data into specially designed 2D color barcodes. To achieve it, the authors developed a new COBRA barcode optimized for streaming between small-size screen and low-speed camera of smartphones.

As presented, a lot of various solutions for barcoding exist, however there still are different relevant problems concerning data barcode representation improvement, which requires some new approaches in data compressing.

3 MATERIALS AND METHODS

A barcode symbol consists of barcode patterns. In its turn, a barcode pattern consists of s elements, which are matrix cells on a carrier. Each cell can be either black, gray, or white.

We assume that maximum capacity of a barcode symbol equals V barcode patterns. In this case, $V \le 3^s$, where 3 is a number of colors and s is a number of cells in the barcode pattern (Fig. 2). As shown in Table 1, $V_{\max} = 3^s$ barcode patterns. A set of all possible barcode patterns with a fixed s forms the multiplicity, or the alphabet Ω of cardinality $P_\Omega = 3^s$. Let us call this alphabet a symbolism of barcode.

The symbolism of barcode consists of informational patterns Ω_{inf} and auxiliary patterns Ω_{aux}, i.e. $\Omega = \Omega_{\text{inf}} \cup \Omega_{\text{aux}}$. Capacity of informational patterns is $P_{\Omega_{\text{inf}}}$ and capacity of auxiliary patterns is $P_{\Omega_{\text{aux}}}$. Thus, $P_{\Omega_{\text{inf}}} + P_{\Omega_{\text{aux}}} = 3^s$.

To represent information on a carrier, we use $P_{\Omega_{\text{inf}}}$ informational barcode patterns.

Auxiliary patterns are used to switch between encoding modes, indicate START and STOP barcode patterns and setup a scanner.

Table 1 – Dependence of barcode pattern maximal capacity on barcode pattern digital capacity

s value	3^s value	Maximal capacity of a barcode pattern, $V_{\max}$	Type of BGW-Code
4	3^4	81	Very small
5	3^5	243	Small
6	3^6	729	Middle
7	3^7	2187	Large
8	3^8	6561	Very large
9	3^9	19683	Ultra large

DOI 10.15588/1607-3274-2020-1-13

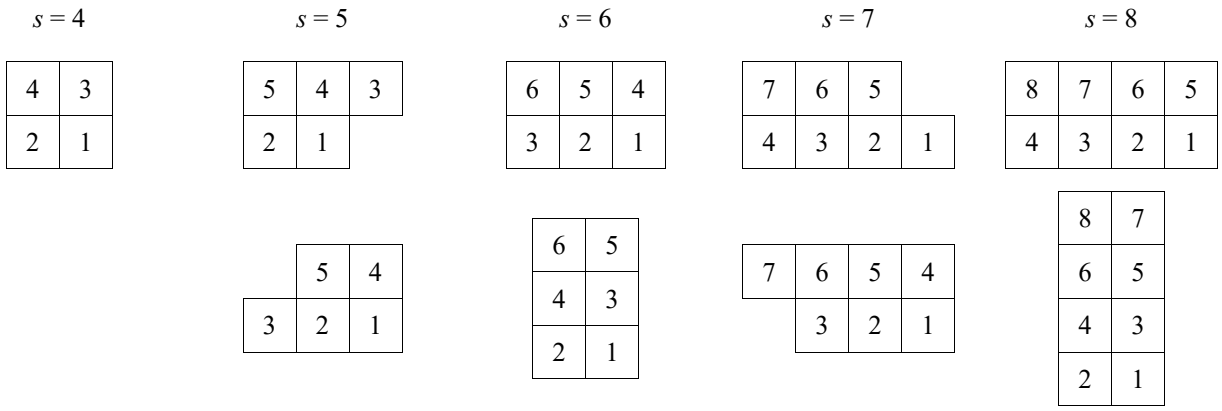


Figure 2 – Bits arrangement in a barcode pattern according to digit capacity *s*

Let us represent an input *h* long alphanumeric sequence $T = t_1 t_2 \dots t_h$. Elements of the sequence belong to extended ASCII, i.e. $t_i \in$ ASCII(256), $i = 1, 2, \dots, h$.

Let the set ASCII be presented as the following: ASCII = $\{L \cup D \cup C\}$ where *L* is a letters set, *D* is a digits set, and *C* is a special symbols set.

The sequence *T* divides into adjacent subsets that are consisted of elements belonged to one of ASCII subsets:

$$T = w_1 w_2 \dots w_k, \quad (1)$$

where $w_i = t_1 t_2 \dots t_n$ is a subsequence of the input sequence which contains elements t_i of only one set of ASCII subsets, namely *L*, *D* or *C*. The subsequences $w_1, w_2, \dots, w_k$ can be situated in any order in the input sequence *T*.

Let alphanumeric symbols t_i belong to an alphabet *A*, which belongs to ASCII, i.e. $t_i \in A$, $A \subset$ ASCII. Cardinality of the alphabet *A* is considered to be equal P_A. The alphabet *A* corresponds with a numeric set $\{0, 1, \dots, P_A - 1\}$ that represents numbers of the symbols of the alphabet *A* in the alphabet.

Now we turn the subsequence $w_i = t_1 t_2 \dots t_n$ formed out of the symbols of the alphabet *A* into a barcode. In the barcode form, the subsequence $t_1 t_2 \dots t_n$ corresponds to a subsequence u_z consisted of *m* barcode patterns: $u_z = \omega_1 \omega_2 \dots \omega_m$ where $\omega_i \subset \Omega_{\text{inf}}$.

In turn, the alphabet Ω_{inf} corresponds with a numeric set $\{0, 1, \dots, P_{\Omega_{\text{inf}}} - 1\}$, as barcode patterns of the barcode symbolism, which are used for representing textual data, can be numbered from 0 to $P_{\Omega_{\text{inf}}}$.

The above-mentioned transformation is considered as $w_i \to u_z$, i.e. $(t_1 t_2 \dots t_n) \to (\omega_1 \omega_2 \dots \omega_m)$. Then the condition for transforming *n* adjacent symbols belonging to the alphabet *A* into *m* barcode patterns belonging to the alphabet Ω_{inf} (i.e. the barcode symbolism) is as follows:

$$n(P_A) \to m\left(P_{\Omega_{\text{inf}}}\right). \quad (2)$$

Practically, the transformation (2) means a transformation of *n*-digits number in a notation P_A into *m*-digits number in a notation $P_{\Omega_{\text{inf}}}$.

The transformation (2) will be with compression if $n \rceil \log_3 P_A [> m \rceil \log_3 P_{\Omega_{\text{inf}}} [$ and at the same time, $P_A^n - 1 \le P_{\Omega_{\text{inf}}}^m - 1$ where $P_A^n - 1$ and $P_{\Omega_{\text{inf}}}^m - 1$ are quantitative equivalents of, correspondingly, maximal *n*-digits number in a notation P_A and maximal *m*-digits number in a notation $P_{\Omega_{\text{inf}}}$. $n \rceil \log_3 P_A [$ is a length of the ternary sequence, which corresponds to an alphanumeric sequence $w_i = t_1 t_2 \dots t_n$.

To let *n* symbols long alphanumeric subsequence $w_i = t_1 t_2 \dots t_n$ be transformed into *m* long barcode patterns subsequence $u_z = \omega_1 \omega_2 \dots \omega_m$ with compression, it is necessary that the following condition is met:

$$\begin{cases} P_A^n - 1 \le P_{\Omega_{\text{inf}}}^m - 1, \\ n \rceil \log_3 P_A [> m \rceil \log_3 P_{\Omega_{\text{inf}}} [. \end{cases}$$

It is important that $\rceil \log_3 P_A [= s$. Thus, it is necessary and sufficient condition for transforming the subsequence $w_i = t_1 t_2 \dots t_n$ of alphanumeric symbols from the alphabet *A* into the subsequence $\omega_1 \omega_2 \dots \omega_m$ of barcode patterns from the symbolism Ω_{inf} with compression:

$$\begin{cases} P_A^n \le P_{\Omega_{\text{inf}}}^m; \\ n \rceil \log_3 P_A [> ms, \end{cases} \quad (3)$$

DOI 10.15588/1607-3274-2020-1-13

where ms is a number of tricolor cells on a carrier that represent the subsequence w_i.

Such a transformation is necessary for ensuring space-keeping data representation on a carrier and increasing data density of barcode patterns with their fixed geometrical dimensions and at the same time, unchanging carrier size.

We define a degree of input data compression as a ratio of a length of ternary sequence that corresponds with alphanumeric sequence w_i to a number of cells on a carrier that represents subsequence w_i in barcoded form and refer to it as a compression coefficient:

$$U^{(s)}_{P_{\Omega_{\inf}}}(P_A) = \frac{n\rceil \log_3 P_A[}{ms}. \quad (4)$$

Thus, the data compression problem (3) consists in finding such P_A with fixed $P_{\Omega_{\inf}}$, as well as parameters n and m, so that maximum value of $U^{(s)}_{P_{\Omega_{\inf}}}(P_A)$ is guaranteed.

4 EXPERIMENTS

Let us consider s = 7, i.e. large BGW-Code will be considered, thereby achieving barcode symbols with capacity of 2187 barcode patterns (see Table 1). Thus, the symbolism of the barcode comprises 2187 tricolor barcode patterns, each of which consists of 7 cells. The barcode patterns correspond to a numeric set {0, 1, …, 2186}.

15 barcode patterns let be considered as auxiliary ones. In this case, 2172 barcode patterns remain for representing information.

The inequality system (3) for large BGW-Code is as follows:

$$\begin{cases} P_A^n \le 2172^m \\ n\rceil \log_3 P_A[> 7m \end{cases}. \quad (5)$$

Now we need to solve the system (5) relatively to P_A. Only integer values are considered as solutions, and the compression coefficient is calculated according to (4): $U^{(s)}_{2172}(P_A) = \frac{n\rceil \log_3 P_A[}{7m}$. Such n and m values are sought for each P_A, so as maximum compression coefficient will be achieved. Fig. 3 defines dependence of a compression coefficient on a cardinality P_A.

The Table 2 below shows some integer solutions (P_A, n, m) that provide compression of data when it is represented as a barcode.

If w_i is the subsequence of decimal numbers, maximum compression with a compression coefficient 1.429 will be achieved when each 10-digits subsequence of adjacent decimal numbers corresponds with 3 barcode patterns. Practically, transformation "10" → "3" means that 10-digits decimal number transforms into 3-digits number in a notation 2172.

The alphabet P_A = 28 that is composed of 16 hexadecimal numbers {0 – F} and 12 other random symbols, such as letters and special characters, provides a possibility to represent hexadecimal sequences, e.g. exe-files, in the form of barcode. In this regard, we should use the transformation "16" → "7" where each 16-digits subsequence corresponds with 7 barcode patterns, which means that 16-digits hexadecimal number is transformed into 7-digits number in a notation 2172. With this transformation, 16-digits subsequence is compressed with a compression coefficient 1.306 (see Table 2).

The alphabet P_A = 267 provides a possibility to represent an input information comprised of any ASCII(256) symbols in the form of barcode.

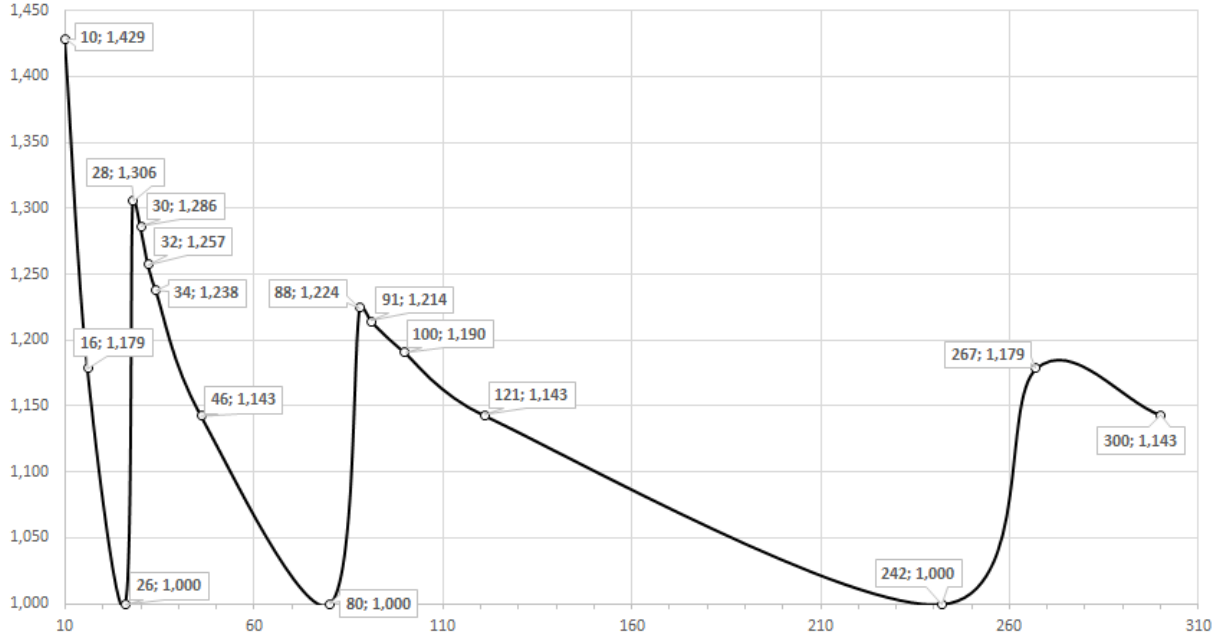


Figure 3 – Dependence of a compression coefficient on a cardinality P_A of the alphabet when s = 7

DOI 10.15588/1607-3274-2020-1-13

In this case, the compression coefficient equals 1.179 and each subsequence consisted of 11 alphanumeric symbols corresponds with 8 barcode patterns, i.e. the transformation "11" → "8" means that 11-digits number in the notation 267 transforms into 8-digits number in the notation 2172.

From Fig. 3 and Table 2, we can assume that it is most appropriate to use the following 4 compression modes for an input alphanumeric sequence T (see Fig. 4):

– DNM, Decimal Numbers Mode: PA = 10, the mode for compressing decimal subsequences,

– HNM, Hexadecimal Numbers Mode: PA = 28, the mode for compressing hexadecimal subsequences,

– TDM, Textual Data Mode: PA = 88, the mode for compressing alphanumeric data (overall number of symbols in the text shall not exceed 88),

– ASM, ASCII Symbols Mode: PA = 267, the mode for compressing subsequences composed of ASCII(256) symbols.

To switch between the modes, switch symbols $\omega_{2172} - \omega_{2181}$ that corresponds to appropriate auxiliary barcode patterns, the mode switchers, are used. For instance, the mode switcher ω_{2172} corresponds to the barcode pattern number 2172 in the barcode symbolism and provides a transition from ASM to DNM.

Before a barcode image be plotting on a carrier, the input alphanumeric sequence (1), which shall be represented as a barcode, is reduced to the following form: $\tilde{T} = \omega w_1 \omega w_2 \ldots \omega w_k$, where ω is a mode switcher, $\omega \in \{\omega_{2172}, \ldots, \omega_{2181}\}$. Each subsequence $w_i = t_1 t_2 \ldots t_n$ is encoded under the rules of the appropriate mode.

In DNM, the following transformation is performed: $\sum_{i=1}^{10} t_i 10^{i-1} \to \sum_{r=1}^{3} \omega_r 2172^r$ where $t_i \in \{0, 1, 2, \ldots, 9\}$.

In HNM, the transformation $\sum_{i=1}^{16} t_i 28^{i-1} \to \sum_{r=1}^{7} \omega_r 2172^r$ is performed, where $t_i \in \{0, 1, 2, \ldots, 27\}$.

In TDM, the transformation $\sum_{i=1}^{12} t_i 88^{i-1} \to \sum_{r=1}^{7} \omega_r 2172^r$ is performed, where $t_i \in \{0, 1, 2, \ldots, 87\}$.

In ASM, the following transformation is performed: $\sum_{i=1}^{11} t_i 267^{i-1} \to \sum_{r=1}^{8} \omega_r 2172^r$ where $t_i \in \{0, 1, 2, \ldots, 266\}$.

As a result of these transformations, the array of number from the range $0 \div 2171$ is obtained instead of the input alphanumeric sequence T. Then, each number is replaced by the appropriate barcode pattern from the symbolism and is arranged on a carrier. The barcode symbol can have either square or rectangle shape.

Table 2 – Some integer solutions of the system (5) for $s = 7$

Cardinality of alphabet A, P_A	Type of transformation, $n \to m$	Compression coefficient, $U^{(7)}_{2172}(P_A)$	What can be represented
10	10 → 3	1.429	Decimal numeric sequences
28	16 → 7	1.306	Hexadecimal numeric sequences
30	**9 → 4**	**1.286**	Textual information comprised of uppercase Latin-script letters
32	11 → 5	1.257	
34	13 → 6	1.238	
88	**12 → 7**	**1.224**	Textual information comprised of Latin alphabet and/or Cyrillic alphabet
91	17 → 10	1.214	
100	5 → 3	1.190	Shortened ASCII(128)
267	11 → 8	1.179	ASCII(256)

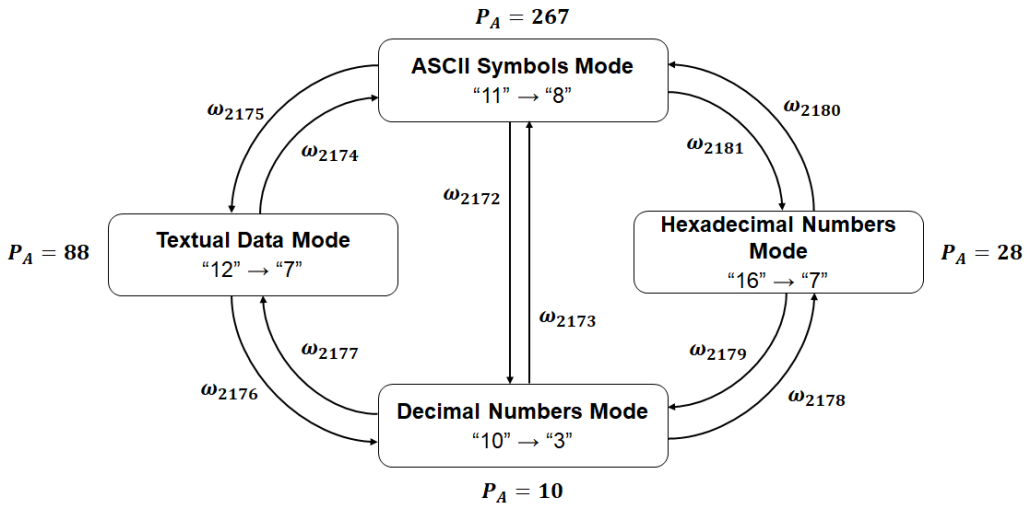


Figure 4 – Interconnection between the compression modes

DOI 10.15588/1607-3274-2020-1-13

The method of compression of alphanumeric information, which is to be represented as a BGW-Code, is as follows:

1. Taking into account the parameter V, which is the capacity of a barcode symbol, the cardinality of the symbolism Ω is defined as $P_\Omega = 3^s$ where $s = \rceil \log_3 V \lceil$.

2. A necessary number $P_{\Omega_{\text{aux}}}$ of auxiliary barcode patterns shall be chosen. A number of informational barcode patterns is $P_{\Omega_{\text{inf}}} = P_\Omega - P_{\Omega_{\text{aux}}}$.

3. Local extrema of a compression coefficient $U^{(s)}_{P_{\Omega_{\text{inf}}}}$ are found. For this purpose, the system (3) is solved with established s and $P_{\Omega_{\text{inf}}}$. Solutions are a set of alphabet cardinality P_{A_i} values, under which a local extremum of a compression coefficient $U^{(s)}_{P_{\Omega_{\text{inf}}}}$ is achieved, and a type of transformation "n" → "m" for each P_{A_i}.

4. A number ψ of the compression modes is determined, where ψ is a number of chosen local extrema of the compression coefficient $U^{(s)}_{P_{\Omega_{\text{inf}}}}$. Modes transmission rules shall be defined.

5. An appropriate alphabet is formed for each mode.

6. Rules of partitioning input alphanumeric sequences to subsequences of adjacent symbols, which consists only of symbols of the appropriate mode alphabet, are formed.

7. Each obtained subsequence is processed by the rules of the appropriate mode and transformed into numeric form, which is a sequence of numbers from the range $0 \div P_{\Omega_{\text{inf}}} - 1$.

The proposed method of data compression can be used for random symbol sequences input using a keyboard.

5 RESULTS

Let us consider an input textual information consisted of 82 ASCII symbols, which shall be represented in the form of BGW-Code (all the data are fictional):

OLEKSIY KLYMENKO, 01/02/1990,
XX83629, 36 KHRESHCHATYK STR, (6)
0156318044 KYIV-21 UKR.

As a result of the analysis of the alphanumeric sequence (6), which is conducted by the appropriate software, the syntax analyzer, the following string is obtained:

12 12 12
OLEKSIY_KLYMENKO,_01/02/1990,_XX8362
12 12 10 12
9,_36_KHRESHCHATYK_STR,_ω_{2176}0156318044ω_{2177}_KYIV-21_UKR.

In this string, 2 mode switchers were inserted: ω_{2176}, to switch from TDM to DNM, and ω_{2177}, to switch from DNM to TDM.

Each 12-digits subsequence (there are 6 such subsequences) shall be replaced by 7 numbers (the transformation "12" → "7") from the range $0 \div 2171$, and each 10-digits subsequence consisted of decimal numbers shall be replaced by 3 numbers (the transformation "10" → "3") from the same range $0 \div 2171$.

Eventually, a numeric sequence comprised of 47 numbers from the range $0 \div 2186$ (as 2172 informational barcode patterns with numbers $0 \div 2171$ and 15 auxiliary barcode patterns with numbers $2171 \div 2186$ compose the symbolism Ω), including 2 mode switchers ω_{2176} and ω_{2177}, which correspond to numbers 2176 and 2177, is obtained as follows:

0202221 2111100 1212101 1012020 1221120 0200011
2021120 0110022 2212011 0010200 2100120 0110020
1000111 2102222 0000002 1201022 1100022 1001010
0020101 1010022 2200200 0000211 2221122 1012121
2002011 1202111 2022222 2022212 0101121 1210121
2020021 2220112 0210210 2002012 0122212 2222121
0001020 0101212 1212222 2222122 0012101 0210100
1002002 1212012 1202122 0212202 1100201.

To obtain a barcode symbol, each of 47 symbols shall correspond to a barcode pattern consisted of 7 tricolor cells (see Fig. 2).

Since a barcode symbol acquires the rectangle shape, one more barcode pattern shall be added to 47 patterns of the barcode: ω_{2182} that represents Pad symbol, a placeholder.

Thus, the barcode symbol presented in Fig. 1 comprised of 336 tricolor cells, as 7×48 barcode patterns is equal to 336. The dimension of the barcode symbol is 16×21 cells.

6 DISCUSSION

Let us consider the obtained results in order to discuss efficiency of the method.

If the textual sequence (6) consisted of 82 symbols of the alphabet with cardinality $P_A = 88$ is represented on a carrier as a black-and-white barcode image, it would require $82 \rceil \log_2 88 \lceil = 574$ black-and-white cells. If the same sequence (6) is represented as a tricolor BGW image, it would require $82 \rceil \log_2 88 \lceil = 410$ tricolor cells. In other words, data density of the barcode symbol increases approximately in 574/410 = 1.4 times. It happens due to transition from two-color to tricolor image.

As a result of the use of both three colors and the proposed compression method, it takes 336 tricolor cells (see Fig. 1) to represent the textual sequence (6).

Thus, the proposed method provides data density with the compression coefficient 410/336 = 1.22. The total effect of the transition from two-color to tricolor image alongside using the compression method provides compression with the coefficient 574/336 = 1.708.

Increasing data density by 1.708 times is assured due to trichromatism (1.4) and the compression method (1.22). Indeed, $1.4 \times 1.22 = 1.708$.

DOI 10.15588/1607-3274-2020-1-13

Thus, the multicolor barcoding method proposed in the paper allows to perceptibly increase amounts of information that can be stored in the form of barcode. In the example above, the data density effect is up to 70%. Depending on the area of application and a specific use case, it can ensure significant benefits, such as autonomous access to large amount of actual data, instead of keeping some general information with a link to more data, which is much more convenient, reliable, and in some cases, even more secure way to get information.

CONCLUSIONS

As barcodes are widely used in multiple fields of human activity, there still are various issues concerned with encoding information. And one of such problems is barcoding more data using the same area of barcode graphical representation.

The scientific novelty of obtained results is that the method of tricolor barcoding with compression is firstly proposed. Compression is achieved due to input data transformation into barcode patterns. The proposed transformation method provides transforming a subsequence of input characters into a shorter subsequence of barcode patterns which will form then the resulting barcode symbol. Use of three colors in the barcode ensures additional compression due to use of ternary notation. Combination of these two approaches allows to barcode more information using the same area of a barcode symbol then it would be with use of binary notation without compression.

The practical significance of the proposed method is that more textual information can be encoded in the form of a single barcode symbol. It can be successfully used in various practical applications when size of the overall barcode symbol is essential, especially when there are quite a lot of data to be barcoded.

Prospects for further research are to study noise immunity issue, which must be considered in order to guarantee error-free scanning and increase reliability of the barcode, and to extend the barcoding software to be used in any language, not only Latin and Cyrillic alphabets but also some specific alphabets, such as Korean, Georgian, Arabic etc., and hieroglyphics.

REFERENCES

1. Forrest P. J., Campbell M. J., Fullerton T. J., Celentano M. J., Brewer R. K. U.S. Patent 6,049,781. Relocation tracking system and method /; applicant HP Enterprise Services LLC. No. US08/634,479 ; appdate 18.04.1996 ; pubdate 11.04.2000.
2. Sriram Th., et al. Applications of barcode technology in automated storage and retrieval systems, *Industrial Electronics, Control, and Instrumentation : The 22nd International Conference, 9 August 1996 : proceedings*. Taipei, IEEE IECON, 1996, Vol. 1.
3. Sun H.-Y. The application of barcode technology in logistics and warehouse management, *Education Technology and Computer Science : The First International Workshop, 7–8 March 2009 : proceedings*. Wuhan, IEEE, 2009, Vol. 3.
4. Kaminsky M. A., Choi J., Lim S., Palmer M. C. U.S. Patent USD710362S1. Barcode scanning device; applicant Motorola Solutions Inc. No. US29/458,380 ; appdate 19.06.2013 ; pubdate 05.08.2014.
5. Bridgelall R., Katz J., Goren D., Dvorkis P., Li Y. U.S. Patent US5988508A. Laser scanning system and scanning method for reading 1-D and 2-D barcode symbols /; applicant Symbol Technologies LLC. No. US08/871,615 ; appdate 10.06.1997 ; pubdate 23.11.1999.
6. Grillo A., Lentini A. et al. High Capacity Colored Two Dimensional Codes, *Computer Science and Information Technology : The International Multiconference, 18–20 October 2010, proceedings*. Wisla, IEEE, 2010, pp. 709–716.
7. High capacity color barcodes [Electronic resource]. Access mode: http://research.microsoft.com/en-us/projects/hccb/.
8. Cattrone P. U.S. Patent US7478746B2. Two-dimensional color barcode and method of generating and decoding the same /; applicant Konica Minolta Laboratory USA Inc. No. US11/444,288 ; appdate 31.05.2006 ; pubdate 06.12.2007.
9. Barrus J., Wolff G. J.; U.S. Patent US7150399B2. Embedding barcode data in an auxiliary field of an image file / applicant Ricoh Co Ltd. No. US10/865,584 ; appdate 09.06.2004 ; pubdate 15.12.2005.
10. Bagherinia H., Manduchi R. A theory of color barcodes, *Computer Vision Workshops (ICCV Workshops) : International Conference, 6–13 November 2011 : proceedings*. Barcelona, IEEE, 2011.
11. Querini M., Grillo A. et al. 2D Color Barcodes for Mobile Phones, *International Journal of Computer Science and Applications*, 2011, Vol. 8, No. 1, pp. 135–155.
12. Bulan O., Monga V., Sharma G. High capacity color barcodes using dot orientation and color separability, *Media Forensics and Security : Symposium, 19–21 January 2009 : proceedings*. San Jose, SPIE, 2009, Vol. 7254.
13. Parikh D., Jancke G. Localization and segmentation of a 2D high capacity color barcode, *Applications of Computer Vision : The IEEE Workshop, 7–9 January 2008 : proceedings*. Copper Mountain, IEEE, 2008.
14. Bulan O., Monga V., Sharma G. U.S. Patent US8100330B2. Method for encoding and decoding data in a color barcode pattern; applicant Xerox Corp. No. US12/436,456 ; appdate 06.05.2009 ; pubdate 11.11.2010.
15. Hao T., Zhou R., Xing G. COBRA: Color barcode streaming for smartphone systems, *Mobile systems, applications, and services (MobiSys '12) : The 10th International Conference, 25–29 June 2012 : proceedings*. Low Wood Bay, ACM, 2012, pp. 85–98.

Received 27.06.2019.
Accepted 05.02.2020.

УДК 004.627

УЩІЛЬНЕННЯ ДАНИХ ПРИ ЧОРНО-СІРО-БІЛОМУ ШТРИХОВОМУ КОДУВАННІ ДАНИХ

Дичка І. А. – д-р техн. наук, професор, декан факультету прикладної математики Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.

DOI 10.15588/1607-3274-2020-1-13

Онай М. В. – канд. техн. наук, доцент кафедри програмного забезпечення комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.

Сулема О. К. – аспірантка кафедри програмного забезпечення комп'ютерних систем Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.

АНОТАЦІЯ

Актуальність. Розглянуто задачу отримання триколірних матричних штрихових кодів підвищеної інформаційної щільності для представлення текстової інформації. Запропонований метод базується на структурно-логічному підході та використанні трьох кольорів у штрихкодовій позначці замість двох, які широко використовуються у чорно-білих штрихових кодах. Такий підхід дозволяє збільшити інформаційну щільність, зберігаючи ту саму площу, що мав би чорно-білий штриховий код.

Метод. Запропоновано метод триколірного матричного штрихового кодування з ущільненням. Основна ідея методу полягає у ущільненні вхідної текстової інформації на етапі перетворення алфавітно-цифрової послідовності у сукупність штрихкодових знаків, які сформують результуючу штрихкодову позначку. Це є можливим завдяки проміжному перетворенню вхідних символів з початкової системи числення, яка визначається потужністю вхідного алфавіту, у систему числення, визначену потужністю алфавіту штрихкодових знаків. Від вибору початкового алфавіту залежить остаточне ущільнення, тому обрання алфавітів для текстової інформації, яка кодується, є важливим кроком пропонованого методу. Використання трьох кольорів (чорного, сірого та білого) замість двох (чорного та білого) також відіграє значну роль у процесі створення штрихкодової позначки з підвищеною інформаційною щільністю. Оскільки використовується трійкова система числення, друге перетворення вхідної послідовності з проміжної системи числення у трійкову систему числення забезпечує ще більше ущільнення. Запропонований метод дозволяє представляти більше текстової інформації у вигляді однієї штрихкодової позначки, ніж можна представити, використовуючи підходи двоколірного штрихового кодування.

Результати. Розроблено та описано метод триколірного матричного штрихового кодування з ущільненням. Наведено приклад практичного застосування методу на наборі тестових даних із демонстрацією відповідної триколірної матричної штрихкодової позначки.

Висновки. Проведені в рамках цього дослідження експерименти підтвердили, що пропонований метод забезпечує більшу інформаційну щільність порівняно з чорно-білими матричними штриховими кодами. Перспективи подальших досліджень можуть полягати у вивченні питання завадостійкості, яка необхідна для забезпечення безпомилкового сканування та підвищеної надійності штрихового коду, а також у розширенні програмного забезпечення для штрихового кодування на інші мови та нетипові алфавіти.

КЛЮЧОВІ СЛОВА: захист текстової інформації, автоматичне введення даних, штрихове кодування, матричні штрихові коди, триколірні штрихові коди, штрихові коди у градієнті сірого, чорно-сіро-білі штрихові коди.

УДК 004.627

СЖАТИЕ ДАННЫХ ПРИ ЧЕРНО-СЕРО-БЕЛОМ ШТРИХОВОМ КОДИРОВАНИИ

Дичка И. А. – д-р техн. наук, профессор, декан факультета прикладной математики Национального технического университета Украины «Киевский политехнический институт имени Игоря Сикорского», Киев, Украина.

Онай Н. В. – канд. техн. наук, доцент кафедры программного обеспечения компьютерных систем Национального технического университета Украины «Киевский политехнический институт имени Игоря Сикорского», Киев, Украина.

Сулема О. К. – аспирант кафедры программного обеспечения компьютерных систем Национального технического университета Украины «Киевский политехнический институт имени Игоря Сикорского», Киев, Украина.

АННОТАЦИЯ

Актуальность. Рассмотрена задача получения трехцветных матричных штриховых кодов с повышенной информационной плотностью для представления текстовой информации. Предложенный метод основывается на структурно-логическом подходе и на использовании трех цветов в штрихкодовом изображении вместо двух, которые широко используются в черно-белых штриховых кодах. Такой подход позволяет увеличить информационную плотность, сохраняя ту же самую площадь, которую занимал бы черно-белый штриховой код.

Метод. Предложен метод трехцветного матричного штрихового кодирования со сжатием. Основная идея метода заключается в сжатии введенной текстовой информации на этапе превращения алфавитно-цифровой последовательности в совокупность штрихкодовых знаков, которые формируют конечное штрихкодовое изображение. Это возможно благодаря промежуточному преобразованию входящих символов из исходной системы исчисления, которая определяется мощностью входящего алфавита, в систему исчисления, определенную мощностью алфавита штрихкодовых знаков. От выбора исходного алфавита зависит окончательное сжатие, поэтому выбор алфавитов для кодируемой текстовой информации является важным шагом предложенного метода. Использование трех цветов (черного, серого и белого) вместо двух (черного и белого) также играет немаловажную роль в процессе формирования штрихкодового изображения с повышенной информационной плотностью. Поскольку используется троичная система исчисления, второе преобразование входящей последовательности из промежуточной системы исчисления в троичную систему исчисления обеспечивает еще большее сжатие. Предложенный метод позволяет представлять большое текстовой информации в виде одного штрихкодового изображения, чем можно представить, используя подходы двухцветного штрихового кодирования.

Результаты. Разработан и описан метод трехцветного матричного штрихового кодирования со сжатием. Приведен пример практического использования метода на наборе тестовых данных с демонстрацией соответствующего трехцветного матричного штрихкодового изображения.

DOI 10.15588/1607-3274-2020-1-13

Выводы. Проведенные в рамках этого исследования эксперименты подтвердили, что предложенный метод обеспечивает большую информационную плотность по сравнению с черно-белыми матричными штриховыми кодами. Перспективы дальнейших исследований могут заключаться в изучении вопроса помехоустойчивости, которая необходима для обеспечения безошибочного сканирования и повышенной надежности штрихового кода, а также в расширении программного обеспечнгия для штрихового кодирования на другие языки и нетипичные алфавиты.

КЛЮЧЕВЫЕ СЛОВА: защита текстовой информации, автоматический сбор данных, штриховое кодирование, матричные штриховые коды, трехцветные штриховые коды, штриховые коды в градиенте серого, черно-серо-белые штриховые коды.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Патент США 6,049,781. Relocation tracking system and method / P. J. Forrest, M. J. Campbell, T. J. Fullerton, M. J. Celentano, R. K. Brewer ; заявник HP Enterprise Services LLC. – № US08/634,479 ; заявл. 18.04.1996 ; опубл. 11.04.2000.
2. Applications of barcode technology in automated storage and retrieval systems / [Th. Sriram, et al.] // Industrial Electronics, Control, and Instrumentation : XXII міжнародна конференція, 9 серпня 1996 р. : тези доповідей. – Тайбей : IEEE IECON, 1996. – Т. 1.
3. Sun H.-Y. The application of barcode technology in logistics and warehouse management / H.-Y. Sun // Education Technology and Computer Science : I-й міжнародний семінар, 7–8 березня 2009 р. : тези доповідей. – Ухань : IEEE, 2009. – Т. 3.
4. Патент США USD710362S1. Barcode scanning device / M. A. Kaminsky, J. Choi, S. Lim, M. C. Palmer ; заявник Motorola Solutions Inc. – № US29/458,380 ; заявл. 19.06.2013 ; опубл. 05.08.2014.
5. Патент США US5988508A. Laser scanning system and scanning method for reading 1-D and 2-D barcode symbols / R. Bridgelall, J. Katz, D. Goren, P. Dvorkis, Y. Li ; заявник Symbol Technologies LLC. – Application US08/871,615 ; заявл. 10.06.1997 ; опубл. 23.11.1999.
6. High Capacity Colored Two Dimensional Codes / [A. Grillo, A. Lentini, et al.] // Computer Science and Information Technology : Міжнародна мультиконференція, 18–20 жовтня 2010 р. : тези доповідей. – Вісла : IEEE, 2010. – С. 709–716.
7. High capacity color barcodes [Електрон. ресурс]. – Режим доступу: http://research.microsoft.com/en-us/projects/hccb/.
8. Патент США US7478746B2. Two-dimensional color barcode and method of generating and decoding the same / P. Cattrone ; заявник Konica Minolta Laboratory USA Inc. – № US11/444,288 ; заявл. 31.05.2006 ; опубл. 06.12.2007.
9. Патент США US7150399B2. Embedding barcode data in an auxiliary field of an image file / J. Barrus, G. J. Wolff ; заявник Ricoh Co Ltd. – № US10/865,584 ; заявл. 09.06.2004 ; опубл. 15.12.2005.
10. Bagherinia H. A theory of color barcodes / H. Bagherinia, R. Manduchi // Computer Vision Workshops (ICCV Workshops) : Міжнародна конференція, 6–13 листопада 2011 р. : тези доповідей. – Барселона : IEEE, 2011.
11. 2D Color Barcodes for Mobile Phones / [M. Querini, A. Grillo, et al.] // International Journal of Computer Science and Applications. – 2011. – Т. 8, № 1. – С. 135–155.
12. Bulan O. High capacity color barcodes using dot orientation and color separability / O. Bulan, V. Monga, G. Sharma // Media Forensics and Security : Симпозіум, 19–21 січня 2009 р. : тези доповідей. – Сан-Хосе : SPIE, 2009. – Т. 7254.
13. Parikh D. Localization and segmentation of a 2D high capacity color barcode / D. Parikh, G. Jancke // The IEEE Workshop on Applications of Computer Vision : Міжнародний семінар, 7–9 січня 2008 р. : тези доповідей. – Коппер-Маунтен : IEEE, 2008.
14. Патент США US8100330B2. Method for encoding and decoding data in a color barcode pattern / O. Bulan, V. Monga, G. Sharma ; заявник Xerox Corp. – № US12/436,456 ; заявл. 06.05.2009 ; опубл. 11.11.2010.
15. Hao T. COBRA: Color barcode streaming for smartphone systems / T. Hao, R. Zhou, G. Xing // Mobile systems, applications, and services (MobiSys'12) : XX Міжнародна конференція, 25–29 червня 2012 р. : тези доповідей. – Лоу Вуд Бей : ACM, 2012. – С. 85–98.

DOI 10.15588/1607-3274-2020-1-13

UDC 004.02

DECISION SUPPORT TECHNOLOGY FOR SPRINT PLANNING

Melnyk K. V. – PhD, Associate Professor of the Department of Software Engineering and Management Information Technologies, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine.

Hlushko V. N. – Senior Consultant, Solution Architect, GlobalLogic, Kharkiv, Ukraine.

Borysova N. V. – PhD, Associate Professor of the Department of Intelligent Computer Systems, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine.

ABSTRACT

Context. The article describes the relevant planning process of software projects, planning problems and different solutions to these problems basis on use of the Scrum methodology.

Objective. The purpose of the work is to develop the technology for solving the sprint planning task in the face of uncertainty and possible risks from software development standpoint.

Method. The most used software life cycle models are described. The choice of the Scrum as a widely used representative of agile methodology for software development is justified. An analytical review of the methods for estimation of the complexity of user stories is carried out. The major problems of sprint planning are highlighted. The model of the business process to implement an IT-project by Scrum in the form of an BPMN-diagram has been developed. The algorithm to solve the problem of Sprint Backlog planning with uncertainty has been elaborated. The common process of user stories selection from Product Backlog to Sprint Backlog and ways of solving the possible problems are considered. The task of estimation of labor intensity of user stories and the task of risk evaluation in planning are formalized. The technology of user story selection for Sprint Backlog has been developed. Numerical studies of the decision support technology proposed in the article are carried out. It allows suggesting it as the practical tool during sprint planning. The method of adequacy evaluation of proposed technology is offered. The set of key performance indicators for assessing the team performance is selected.

Results. The sprint planning technology was developed, which project managers, product owners and development teams for increasing the effectiveness of decision-making process can use.

Conclusions. The conducted experiments have confirmed the importance of the proposed decision support technology and allow recommending it for use in practice for planning of software projects. Scientific novelty is to improve the sprint planning process with the assistance of the proposed technology, which alleviates uncertainty while defining labor intensity of user stories and decreases time spent on decision making.

KEYWORDS: sprint backlog, planning, uncertainty, labor intensity, user story, user story, selection task, team performance.

ABBREVIATIONS

IS is an Information System;
IT is an Information Technologies;
KPI is a Key Performance Indicators;
LI is a Labor Intensity;
RMS is a Residual Mean Square;
SP is a Story Points;
US is a User Story.

NOMENCLATURE

I is a set of user stories in a sprint;

$|I|$ is a strength of the set I;

$A^{n\times n}$ is a judgment matrix;

a_{ij} is a paired comparison of i-th US and j-th US $(i, j \in I)$;

$s_i, i \in I$ is a labor intensity of i-th US;

g_i is an average relative size of labor intensity of i-th US;

D is a sample variance;

σ_i is the estimation of RMS of i-th US;

CI_i is a confident interval of LI of i-th US;

V is a team velocity;

p_i is a priority of i-th US;

x_i is a selection indicator of i-th US for implementation in the sprint;

TP is a team performance;

w_q is a weight coefficient of q-th key performance indicator;

u_q is an utility function;

KPI_q is a value of q-th key performance indicator;

KPI_q^{worst} is a worst value from set of values of q-th key performance indicator;

KPI_q^{best} is a best value from set of values of q-th key performance indicator.

INTRODUCTION

The attractiveness of Ukraine in the world market of software development services for foreign companies is constantly growing. The IT share in GDP of Ukraine is 4% by the start of 2019 and it is raising [1]. The number of IT companies, complexity of IT projects, requirements for quality and skills of specialists are increasing. As a result, the software development process is making more and more complex. On the one hand, this process is characterized by the complexity of coordination of IT-professionals, where each member of the team has different experience and qualifications. On the other hand, it is necessary to take into account a large number of require-

DOI 10.15588/1607-3274-2020-1-14

ments from users of future programs, which are sometimes controversial. There are many models of software development. The most commonly used software development lifecycle models to date are [2]: Waterfall model, V-model, Incremental iteration model, Prototype model, Spiral model.

All aforementioned models have their advantages and disadvantages for software development. However, Agile methodologies for software development have been getting more and more popularity today. The most utilized among Agile methodologies is the Scrum methodology [3, 4]. It is a flexible development cycle model that allows developers to take advantage of existing coding practices and enables the client to make changes to requirements at any time limiting the negative impact on development teams during the sprint course. Its main feature is involvement of all participants into the process: both client and performer. The use of Scrum allows you to detect and eliminate deviations from the desired result in the earlier stages of software development.

Software development with Scrum consists of small iterations, or sprints, which are essentially small projects. Sprint duration is a fixed time period of 1–4 weeks. It has the same length until the end of the project. When sprint is over, a new working version of the product should be received. The following actions are analysis and refocusing on the new tasks of the next cycle. The effectiveness of sprint and an IT-project, in general, is directly dependent on the planning process, so solving of the sprint planning task is the very important and actual problem nowadays.

The object of study is the planning process of sprint.

The subject of the study is the theoretical and methodological tool for assessing and selection the set of tasks for sprint.

The purpose of this work is to develop a decision support technology for solving the sprint planning task.

1 PROBLEM STATEMENT

The sprint planning activity is a selection of a set of user stories or tasks which development team commits to solve within a single sprint, assessing their complexity and efforts with evaluation of possible risks that may occur while developing software during the sprint. In its turn, it means that the development team literally finds the optimal solution of the planning problem in the face of uncertainty.

The mathematical formulation of the planning task can be presented in the following way:

– to estimate labor intensity of every US $s_i, i \in I$ from the proposed set of user stories;

– to choose subset of user stories $\{US_i\}$ from the proposed set I for next sprint according to labor intensity of every US $s_i, i \in I$ and the team velocity V .

2 REVIEW OF THE LITERATURE

In order to plan a sprint, it is necessary to evaluate the complexity and labor content of user stories. There are most commonly used methods for evaluating the complexity of a story [5, 6]: T-Shirt Sizes method, Planning Poker method, Dot-voting method, Ordering Rule method. Almost all of these methods are based on heuristic approaches. The aforementioned techniques do not need much time, they are quite accurate for comparison of the work efforts of one US to another, and as such they are used on many IT-projects. Regardless on sufficient number of methods for solving the sprint planning problem, they still remain open some issues that development teams face.

To reduce the subjectivity of the judgments of stakeholders, it is necessary to use formal methods based on different mathematical models. There are many scientific articles showed the way of calculating efforts or labor intensity of user stories. The journal publications [7, 8] proposes using of Bayesian network for effort calculation; the reports at scientific conferences [9, 10] demonstrate unusual implementation of Bloom's Taxonomy for computation of complexity of user stories; the works [11, 12] reveal how to construct and use fuzzy logic framework for complexity calculation. These scientific works show the applicability of the proposed methods for assessing the complexity of user stories, but they require a very long period of time for preparation. For each IT-project, the team has to build own model or own framework, and this requires additional financial and time resources, as well as the availability of experienced specialists in the mathematical field. Such a disadvantage can lead to increase project times, which is unacceptable to Product Owner.

All of the aforementioned methods and approaches have a disadvantage: there are no recommendations for planning tasks when the complexity of the estimated user stories exceeds the average sprint velocity for the team. Scrum methodology [3, 4] proposes to decompose large user stories in such cases, and then choose a set of stories, the complexity of which corresponds to the velocity of the sprint. The problem there is that the task of optimal user story selection is task in the face of uncertainty, because the likelihood of different choices is unknown most of the time. In such cases, the team and the Product Owner are guided only by benefits they get in return, with no risk assessment recommendations in every case.

3 MATERIALS AND METHODS

In general, the model of the business process of software development using the Scrum methodology can be represented in the following form (Fig. 1). The Business Process Model and Notation Specification [13] was chosen to model the diagram.

Before the start of software development, Product Owner and development team conducts the first meeting. It is called the Kickoff Meeting, which provides the opportunity for Product Owner to explain vision and scope of the project. The following meeting is the Product Backlog Grooming Meeting, which is devoted to creation of Product Backlog – the main document of the project. Product Backlog can be considered as the software re-

DOI 10.15588/1607-3274-2020-1-14

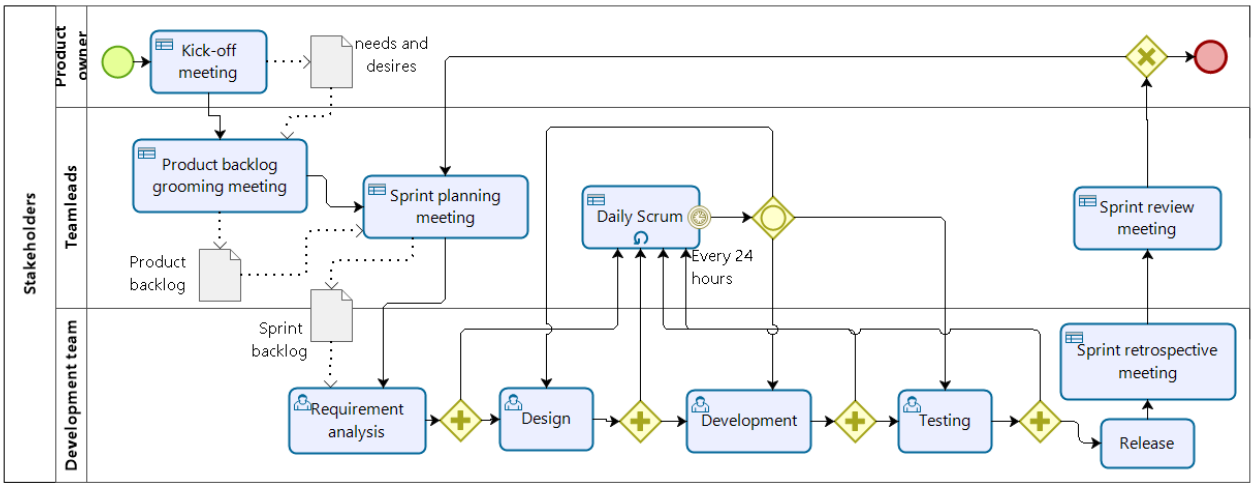


Figure 1 – Scrum model of software development

quirement specification, which consists of epics and user stories. Sprint Planning Meeting is carried out before starting of the sprint. The main purpose is to prioritize and evaluate the content of Product Backlog and form the Sprint Backlog. It means that set of user stories from Product Backlog are selected for the next sprint. The biggest problem at this stage is to correctly evaluate the complexity of each story as well as to assess related uncertainty if any. Every day, a Daily Scrum is conducted to determine the status and progress of work during the sprint, identify early obstacles, and make decisions to change the strategy needed to meet sprint's goals.

The Sprint Retrospective meeting is undertaken on the last day of the sprint. It goes as following:

– team members answer two questions: "what has been done well in the past sprint?" and "what needs to be improved in the next one?";

– highlights improvements of the development process;

– evaluates the efficiency of the team in the past sprint and predicts the expected performance in the next sprint;

– identifies existing problems, proposes possible solutions and assigns team members responsible for them;

– makes estimates of the probability of completion of all necessary work on the product.

Sprint Review Meeting is conducted at the end of the sprint. It may be used by the team to demonstrate the version of the product to all interested stakeholders.

Thus, solving the Sprint planning task allows to increase the effectiveness of decision-making process by project managers. So, the main purpose of this study is to develop decision support technology for sprint planning.

Referring to [3, 4], the general model of solving the sprint planning task can be represented in the following form (Fig. 2).

The first step of planning is to prioritize each user story in Product Backlog. Product Owner handles this activity. Due to the prior prioritization, all user stories are sorted by importance to the business. Typically, Sprint Backlog creation is the selection of user stories with the highest priority from Product Backlog, unless otherwise discussed with the client.

According to the algorithm above, the next step is to solve two problems:

1) estimation of labor intensity of each user story from Sprint Backlog;

2) evaluation of uncertainty and possible risks from software development standpoint.

The determined estimations of Sprint Backlog should be compared with the team velocity. Velocity is a measure of the amount of work the team can do during a single sprint. Depending on the results, there are three options:

1) The labor content of all Sprint Backlog stories is roughly equal to the team's predicted speed and effort. In this case, the Sprint Backlog is not changed, and the team works as usual.

2) If the estimated labor content of Sprint Backlog less than the velocity of the team, then Sprint Backlog is filled with the next user story from the ordered Product Backlog list.

3) A difficult situation arises when the number and complexity of user stories in Sprint Backlog are much greater than the speed of the team. Therefore, the task selection of the highest priority user stories arises, but their number must be as high as possible. Otherwise, Product Owner changes the priorities of user stories or decomposes some of them, and the stories are re-evaluated.

Solving of the problems from above results in creation of Sprint Backlog. After the sprint, the sprint results are analyzed, and the velocity is modified. Its updated value is used for further calculations in the next sprint. Velocity is calculated by totaling the points for all fully completed user stories.

Let's consider the tasks presented above in details.

The task of evaluating uncertainty is to predict implementation of a sprint in the context of incomplete information. There is a risk of failing of team commitments when there is not enough input data for sprint planning. Therefore, one of the important tasks in sprint planning is to assess the uncertainty to mitigate any possible risks that may arise in the following cases:

DOI 10.15588/1607-3274-2020-1-14

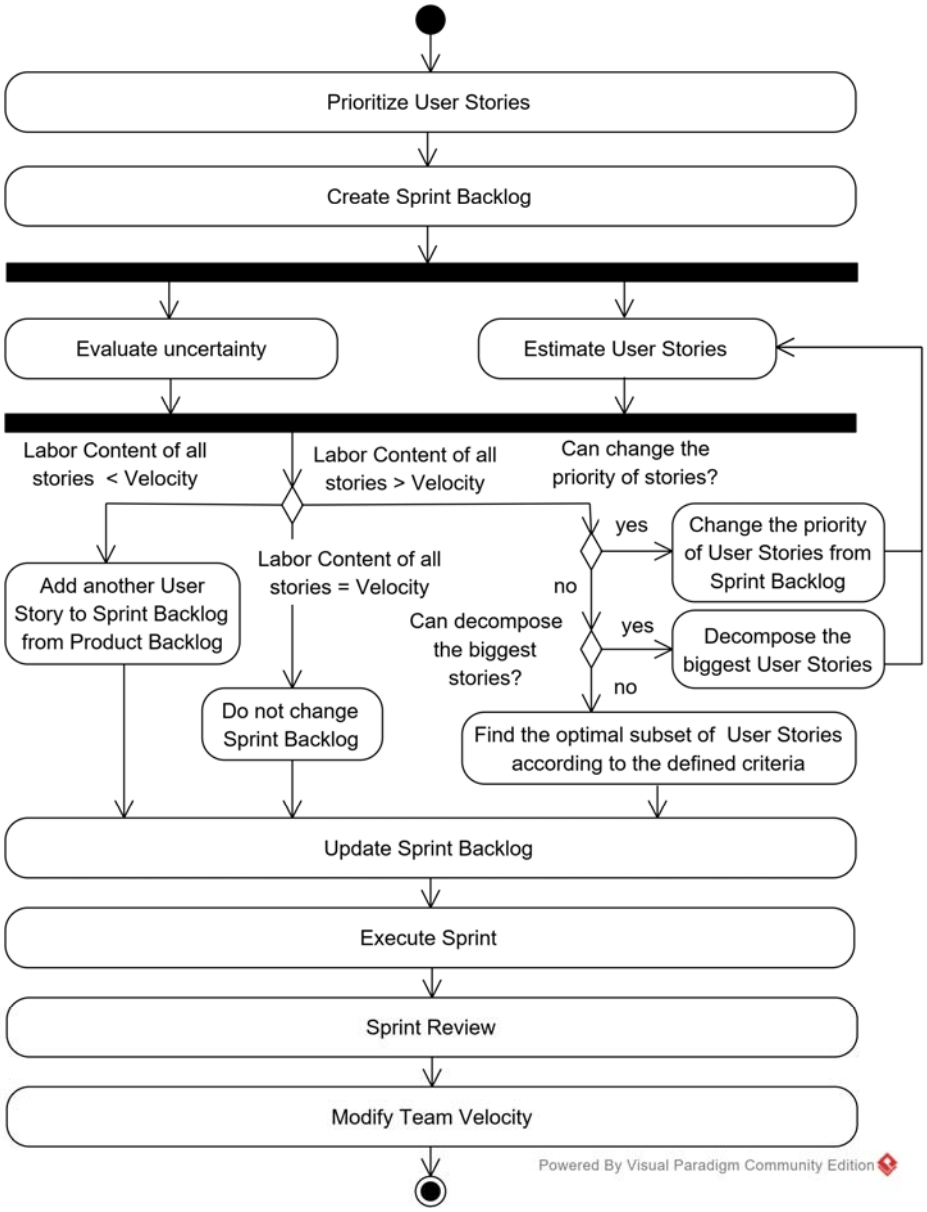


Figure 2 –The algorithm of solving the sprint planning task

1) Changing the team composition:

– temporary occupation of a team member on project tasks unrelated to development, for example, demonstration to a client of a product demo, which may happen even in another country;

– missing employees due to improving their skills or supporting the IT company by attending an IT conference, conducting an online training, involvement into pre-sales activities for new clients, etc.;

– temporary absence of a team member due to the urgent need to use his expertise on other projects;

– temporary absence of a team member due to illness;

– lowering of team member performance due to objective or subjective reasons.

2) Changing of the Product Backlog before the start of the sprint:

– some user stories can be added, deleted or modified on demand of the client, it leads to changing of the labor intensity of these user stories, labor intensity of the sprint and the whole project;

– changing of project scope;

– changing user story priorities in Product Backlog can cause Sprint Backlog failure;

– incorrect estimation of complexity and labor content of the user stories in Sprint Backlog;

– misinterpretation of the client's wishes, in other words, changing of the content of some user stories.

In each case, the Product Owner should decide how to assess any uncertainty and mitigate risks. Uncertainty can be taken into account in two ways.

The first way is to reduce velocity of the next sprint, in case if probability of complications is big enough.

The second way of evaluating of sprint risks is to raise the labor intensity of the user story:

$$LI\ of\ US = f\left(complexity,\ risk,\ efforts\right).$$

DOI 10.15588/1607-3274-2020-1-14

Then the final estimation of the user story should be slightly increased by a couple of units used by the team for the evaluation. However, the adjusted estimates should be used with care in the decision making process as of their expert judgment nature.

We can use different, more formal way of uncertainty evaluation – calculation of interval estimation instead of point estimation:

$$Confident\ interval = \left[LI\ of\ US \pm uncertainty\right]. \quad (1)$$

As a measure of uncertainty, it is suggested to use the residual mean square for estimates of the confidence interval. The RMS shows the average deviation of user story estimation from the mathematical expectation of a set of user stories in Sprint Backlog. Therefore, when Sprint Backlog is being evaluated with a high likelihood of time failure, it is highly recommended to use the pessimistic values of the labor intensity of the user story. It means the usage of the maximum value from the confidence interval.

The RMS can also be used to analyze the estimations: the large value of RMS characterizes the imbalance of the user story estimations. This means that the labor intensity of some user stories is very high. According to the Scrum methodology [3, 4, 14], implementation of user story should not exceed one working day or 12 hours. So, the user story should be divided into separate tasks and re-evaluated.

Consider **the task of estimating the labor intensity of the user story**.

User stories in Sprint Backlog are evaluated in units used by the team: man-hours or story points [5, 6, 14]. Although the Scrum methodology recommends story points as an abstract metric for assessing the labor content of user story, some IT companies use time as a unit of story complexity. In the latter case, there are a certain number of hours to complete the story. A more qualified developer can complete the stories in a part of the allotted time, and then begin to perform the next task or switch to tasks that are not directly related to the sprint goals. Moreover, a developer with no experience can spend extra time for solving a specific problem. Thus, estimations of the user story are not objective in this case. This disadvantage can be minimized by using a different rating scale based on comparisons of labor content to implement user story. It is suggested to use Story Points as a unit of measurement.

1 SP is the unit of labor intensity of the story or the effort of the whole team to implement the simplest requirement or user story.

The number of Story Points to develop the same functionality differs from team to team, but this does not mean that time costs will be different, as each team means its value for 1 SP. Assessing stories in SP makes sense only within the same project and the same development team, because the labor content of the tasks is compared with each other.

Thus, in order to evaluate the complexity of the user story, it is necessary to make subjective paired comparisons on the selected scale. Paired Comparison Method is one of the appropriate decision making tools because of its simplicity and effectiveness. It lets to describe values and compares them to each other. In [15] the scale of comparison for subjectively paired comparisons was proposed:

– equal importance – 1;
– moderate importance – 3;
– strong importance – 5;
– very strong importance – 7;
– extreme importance – 9;
– these marks for intermediate cases – 2, 4, 6, 8.

The effectiveness of this scale has been proven by comparison with many other scales in many applications [16].

The process of pairwise comparison is conducted as following. All user stories need to be compared with each other. The obtained estimates are entered into the judgment matrix. When comparing an element with itself, the ratio equals 1. If the first user story is more important than the second, then an integer from the scale is used, otherwise the inverse value is used. The lower off-diagonal elements are determined by the upper off-diagonal elements. The number of different paired comparisons in a rank-ordering of N objects is $N(N-1)/2$. Then the judgment matrix is used to calculate the estimation of user stories. Most scientific papers uses eigenvector to calculate values of user stories [15]. However, Crawford and William in work [16] showed that the geometric mean vector is computationally easier than eigenvector and statistically preferable to the eigenvector.

Let's formalize the process of user story evaluation based on the mathematical apparatus proposed in [15] and [16].

Let us denote I as the set of user stories in the current sprint, the complexity of which should be evaluated, whereas $|I| = n$ is the strength of the set of user stories. Then a_{ij} is the result of a paired comparison of the i-th and j-th $(i, j \in I)$ user stories, which is written to the judgment matrix $A^{n \times n}$, where $a_{ij} = \frac{1}{a_{ji}}$ and $a_{ii} = 1$.

The average relative size g_i of LI of i-th US is calculated as geometric mean of judgments

$$g_i = \left(\prod_{j=1}^{n} a_{ij} \right)^{\frac{1}{n}}, i \in I. \quad (2)$$

Let us assume $k, k \in I$ is the number of the US chosen as the standard. There is recommendation for the first sprint to choose a standard user story, which labor content is known from team past experience on similar projects, or such story that has the minimal labor content. Standard US for the following sprints is the one which complexity in SP can be found in the easiest way based on the experience of the previous sprints.

Suppose $s_k, k \in I$ is a known value of LI in story points of standard US. According to aforementioned des-

DOI 10.15588/1607-3274-2020-1-14

ignations the labor intensity of i-th US can be calculated by the following formula [17]:

$$s_i = \frac{s_k g_i}{g_k}, i \neq k, i \in I, k \in I, \quad (3)$$

where $g_k, k \in I$ was calculated by the formula (2).

It is necessary to use formula (1) to determine uncertainty of the complexity of user stories, where uncertainty is determined by residual mean square. Analyzing the input data, it is clear that it is impossible to find the RMS, but only RMS estimation by means of the sample variance. The following formula was proposed in science paper [13] for finding the sample variance D_A of the judgment matrix $A^{n\times n}$ for creation of the sprint backlog:

$$D_A = \frac{1}{\frac{n(n-1)}{2} - (n-1)} \sum_{i<j}^{n} \left(\ln a_{ij} - \ln \frac{g_i}{g_j} \right)^2. \quad (4)$$

Sprint backlog is formed in such a way that there is no idleness of the team. So, the implementation of one user story is a process independent from implementation of another user story. Assume that each story contributes equally to the overall variance. In this case, the sample variance of the sprint backlog is the sum $D[s_i]$ of the sample variances of the every user story

$$D_A = \sum_{i=1}^{n} D[s_i] = n\, D[s_i].$$

The estimation of RMS of i-th US is calculated in following way:

$$\sigma_i = \sqrt{D[s_i]} = \sqrt{\frac{D_A}{n}}. \quad (5)$$

Then confident interval CI_i can be found such as:

$$CI_i = \left[s_i \pm s_i \sigma_i \right]. \quad (6)$$

Thus, the algorithm of finding estimates of the labor intensity of user stories has been presented.

Consider the **task of selection of user story** from Sprint Backlog in the case when it is not necessary to change the priorities of a user story.

In general, the scale of priority evaluation may differ from one IT project to another. Product Owner chooses a way to evaluate the user stories herself.

Let's x_i is the variable that indicate whether or not the i-th user story is selected for implementation in the current backlog sprint: $x_i = 1$ when i-th user story is selected and $x_i = 0$ – otherwise. So, the variable can only take two values

$$x_i = \{0,1\}, i \in I. \quad (7)$$

Sprint backlog has to include the highest number of top priority user stories:

$$\sum_{i=1}^{n} p_i x_i \to \max. \quad (8)$$

The labor content of sprint backlog should not exceed the team velocity

$$\sum_{i=1}^{n} s_i x_i \leq V. \quad (9)$$

Based on the aforementioned objective functions and constraints, the model of selection of user stories for the sprint backlog can be defined in the following way: find the set of user stories satisfying the objective function (8) and constraint (9) under condition (7). This task belongs to the class of integer programming problems with Boolean variables.

4 EXPERIMENTS

Let's consider usage of the decision support technology for sprint planning by example. There is IT-project for creating IS. It is known, that team have chosen Scrum as model of software development lifecycle. The team conducted several sprints, so velocity now is equal to $V = 60\,SP$. The product owner for the current sprint have proposed five user stories. He evaluated priority of each user story on the 5-point scale: $p_1 = 4; p_2 = 5; p_3 = 3; p_4 = 5; p_5 = 3$. The team has compared US to each other. The judgment matrix of pairwise comparisons is presented in Table 1.

Table 1 – The judgment matrix

№ US	1	2	3	4	5
1	1,00	3,00	3,00	0,50	2,00
2	0,33	1,00	0,50	0,25	0,50
3	0,33	2,00	1,00	0,33	0,50
4	2,00	4,00	3,00	1,00	2,00
5	0,50	2,00	2,00	0,50	1,00

In order to obtain the numerical values of the labor intensity s_i of the formula (3), the geometric mean of the labor intensity g_i is calculated by using the formula (2). If this is the first sprint, and there is no information about the complexity of the user stories, then it is recommended to take as the standard a story with a minimum geometric mean value, and to take its complexity as 1 SP. Nevertheless, by the statement of the task, it is known that $s_2 = 10\,SP$, therefore, it acts as a standard for this sprint. To determine the risks the sample variance and estimation of RMS for each user story can be found by formulas (4) and (5) accordingly.

To find the interval estimations, which maximum value the product owner can use as a pessimistic estimations of user stories, the limit values have been calculated by formula (6) basing on obtained results.

According to labor complexity of each US and its priority, it is necessary to use the model of selection of user stories (7)–(9). It allows choosing the set of US from proposed sequence of US for current sprint, if summary labor content of proposed sequence of user stories is much greater than the team velocity.

DOI 10.15588/1607-3274-2020-1-14

5 RESULTS

The results of the calculations of the labor intensity of user stories from proposed IT-project for the current sprint are presented in Table 2.

Table 2 – Complexity of user stories

№ US	g_i	s_i	D_A	σ_i	$s_i\sigma_i$	CI_i	
						min	max
1	1,55	33	0,08	0,13	4	29	37
2	0,46	10			1	9	11
3	0,64	13			2	11	15
4	2,17	47			6	41	53
5	1,00	21			3	18	24

The value of estimation of RMS σ_i is low, it means that obtained estimations are well-balanced.

The graphical representation of the confidence intervals of the estimates of user stories is shown on the Fig. 3 based on the obtained results from the Table 2. It can be seen, that for the fourth user story has the largest difference between minimum and maximum estimations. Product owner may use this information to analyze the user stories, for example, to decompose 4th US into sub-stories, thereby reducing uncertainty.

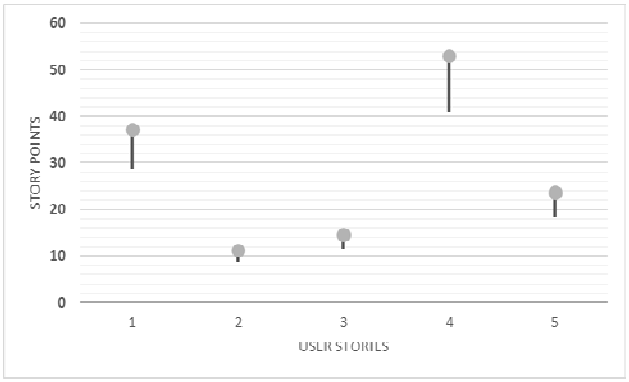


Figure 3 – Interval estimations of user stories

During the previous several sprints, the team showed an average speed of work equal to $V = 60\,SP$. Knowing the complexity of a sprint backlog, one can see that $\sum_{i=1}^{5} s_i = 124\,SP \geq V = 60SP$. Therefore, it is necessary to change the story priorities and reduce the sprint backlog, or to select user stories for the sprint with current priorities. In this particular case, it is not possible to change the priorities according to the conditions of the example, so it is necessary to solve the task of selection of user stories.

Based on priorities and calculated labor intensity of each user story, the objective function (8) can be defined as follows:

$$4x_1 + 5x_2 + 3x_3 + 5x_4 + 3x_5 \to \max .$$

The constraint (9) has the following form:

$$33x_1 + 10x_2 + 13x_3 + 47x_4 + 21x_5 \leq 60 .$$

Taking into account condition (7) the solution of the given problem by simplex method is as following:

$$x_1 = 1; x_2 = 1; x_3 = 1; x_4 = 0; x_5 = 0 .$$

Basing on the calculated results, the user story #4 with high priority has not been selected as the candidate for the sprint backlog. This is because its labor intensity is almost the same as the velocity of the team. In this case, the Product Owner can:

– divide this story into individual cases;
– refine the functionality of this story;
– change priority.

The results for this story are consistent with the recommendations obtained in the evaluation of the complexity of the sprint backlog.

Thus, technology for supporting manager decisions in sprint planning has been considered.

6 DISCUSSION

To assess the adequacy of the developed decision support technology in sprint planning, it is necessary to compare the results of work of the team before applying the technology and after. One of the ways to verify the effectiveness of using the proposed technology in an IT company is to evaluate work efficiency using Performance Management [18]. Performance management is the process of calculating and improving team performance to achieve the goals of the IT-company. To assess team performance, it is necessary to identify set of key performance indicators. We can use the following KPI:

– KPI_1 is a meeting time per sprint;
– KPI_2 is a percentage of missed tasks;
– KPI_3 is the team velocity;
– KPI_4 is a customer satisfaction;
– KPI_5 is a team satisfaction.

To evaluate the team performance TP, the article [18] proposes using of universal mathematical model from [19] as a convolution criterion for KPI:

$$TP = \sqrt{\sum_{q=1}^{5} (w_q u_q)^2} . \qquad (10)$$

The utility function u_q can be found in the following way:

$$u_q = \frac{KPI_q - KPI_q^{worst}}{KPI_q^{best} - KPI_q^{worst}} .$$

Let's consider limit values for proposed KPI. If the team spends more than 2/3 working time on different project meetings, there is the chance the project will be behind the schedule. One of the possible reason is the "analysis paralysis", which means the inability to develop or decide due to overthinking available alternatives and possible outcomes [20]. Another possible reason is "scope creep", it refers to changes, continuous or uncontrolled growth in a project's scope, at any point after the project begins [21]. In the common case, the number of working hours in each sprint is equal to 160 hours, so

DOI 10.15588/1607-3274-2020-1-14

$KPI_1^{worst} \geq 100$ hours. If the $KPI_2^{worst} \geq 60\%$, the project will be behind the schedule as well. The value of team velocity depends on team, project, Product Owner, so best and worst values can be found from set of previous values. The KPI_4 and the KPI_5 should be evaluated by 10-point scale, where 1 is the minimal value for metrics and 10 is the maximum accordingly, so KPI_4^{worst} and KPI_5^{worst} are equal to 1.

To evaluate the team performance, it was considered three sprints: first sprint was conducted without proposed technology, the team worked as usual; during the second and third sprints, the team used decision support technology for sprint planning. The results of the evaluation of the team performance in several sprints calculated by using formula (10) are shown in the Table 4.

Table 4 – Team performance

Key performance indicator	Number of sprint	Value of KPI_q	u_q	w_q	TP
KPI_1	I	28	0,92	0,1	I – 0,26 II – 0,35 III – 0,38
	II	25	0,96		
	III	22	1,00		
KPI_2	I	23	0,76	0,2	
	II	15	0,92		
	III	11	1,00		
KPI_3	I	60	0,00	0,2	
	II	72	0,86		
	III	74	1,00		
KPI_4	I	6	0,56	0,3	
	II	7	0,67		
	III	7	0,67		
KPI_5	I	5	0,44	0,2	
	II	5	0,44		
	III	7	0,67		

The analysis of results of the team performance evaluation from the Table 4 demonstrates the positive dynamics of changes of the team productivity on 9–12%.

The commonly used approaches to decision-making in sprint planning [5–12] allow determining the complexity of user stories in specified units. In comparison to them, the proposed technology allow calculation of the complexity of user stories, takes into account uncertainty of the current sprint, and selects a set of user stories from the sprint backlog when the total complexity of the sprint backlog exceeds the team velocity.

The proposed technology enables increasing of the team productivity and provides additional information for sprint planning. The using of the decision support technology for sprint planning and the obtained results show the feasibility of using the proposed technology in real conditions.

CONCLUSIONS

In the course of this research the decision-making technology for solving the planning problem in the face of uncertainty has been proposed. For this purpose, an analytical review of the methods of estimating the labor intensity of user stories has been conducted. It has revealed the shortcomings of existing approaches. Due to the development of technology of decision support in software development the planning task in the face of uncertainty has been further developed. A sprint planning algorithm has been developed. Formalization of the process of estimation of the labor content of user stories based on the previous experience has been presented. Sprint Backlog reshuffling model in case of extra labor effort needed for its implementation in comparison with team velocity has been developed.

The scientific novelty of the obtained results consists in improvement of the sprint planning process with the assistance of the proposed technology, which helps to reduce uncertainty while defining labor intensity of user stories and Sprint Backlog as a whole. Numerous studies have shown that the use of the proposed technology requires only handy tools, such as Microsoft Excel, OpenOffice Calc, LibreOffice Calc, PlanMaker and others, which do not require from project managers and Product Owners any specific mathematical skills. The results show the **practical significance** of the approach for IT companies and the ability to use the proposed technology in software development projects to increase the effectiveness of decision-making process in uncertainty for project managers, product owners and development teams.

Prospects for further research consists in creating an IS that will reduce the time spent on data processing about US, and will automate the decision-making process for planning the sprint.

ACKNOWLEDGEMENTS

The article is supported by the state budget scientific research project of National Technical University «Kharkiv Polytechnic Institute» of Software Engineering and Management Information Technologies department «Development of models and methods of collecting and automated processing of business information in Web» (state registration number 0119U002556).

REFERENCES

1. Chastka IT-industriyi v ekonomitsi Ukrayiny [Electronic resource]. – Access mode: https://ua.112.ua/suspilstvo/chastka-it-industrii-v-ekonomitsi-ukrainy-stanovyt-4-vvp-kubiv-480452.html
2. Software Engineering Body of Knowledge (SWEBOK) [Electronic resource]. Access mode: https://www.computer.org/education/bodies-of-knowledge/software-engineering
3. A Guide to the Scrum Body of knowledge (SBOKTM Guide) [Electronic resource]. Access mode: http://www.scrumstudy.com/SBOK/SCRUMstudy-SBOK-Guide-2016.pdf

DOI 10.15588/1607-3274-2020-1-14

4. Sutherland J., Sutherland J. J. Scrum: A revolutionary approach to building teams, beating deadlines and boosting productivity. United States, Random House, 2014, 256 p.
5. 7 Agile Estimation Techniques – beyond Planning Poker [Electronic resource]. Access mode: https://technology.amis.nl/2016/03/23/8-agile-estimation-techniques-beyond-planning-poker/
6. Agile Estimation Techniques: A True Estimation in an Agile Project [Electronic resource]. Access mode: https://www.softwaretestinghelp.com/agile-estimation-techniques/
7. Lopez-Martinez J., Ramirez-Noriega A., Juarez-Ramirez R. et al.] User stories complexity estimation using Bayesian networks for inexperienced developers, *Cluster Computing.* 2018, Vol. 21, pp. 715–728. DOI: 10.1007/s10586-017-0996-z
8. Karna H., Gotovac S. Estimating software development effort using Bayesian networks, *Telecommunications and Computer Networks (SoftCOM) : 23rd International Conference on Software, Split, 16–18 September 2015 : proceedings. Split: IEEE*, 2015, pp. 229–233. DOI: 10.1109/SOFTCOM.2015.7314091
9. [Barrera F. E., García M. A., González H. G. et al. Agile Evaluation of the Complexity of User Stories Using the Bloom's Taxonomy, *Computational Science and Computational Intelligence (CSCI'17) : 4th Annual International Conference, Las Vegas, 14–16 December 2017 : proceedings.* Las Vegas, Nevada: Conference Publishing Services (CPS), 2017, pp. 1047–1050. DOI:10.1109/csci.2017.182
10. Castillo-Barrera F. E., Amador-García M., Pérez-González H. G. et al. Adapting Bloom's Taxonomy for an Agile Classification of the Complexity of the User Stories in SCRUM, *Software Engineering Research and Innovation (CONISOFT'18) : 6th International Conference*, San Luis Potosi, 24–26 October 2018, proceedings. San Luis Potosi, IEEE, 2018, pp. 139–145. DOI: 10.1109/CONISOFT.2018.8645899
11. Nassif A. B., Capretz L. F. , Ho D. Estimating software effort based on use case point model using Sugeno Fuzzy inference system, *International Conference on Tools with Artificial Intelligence (ICTAI 2011) : 23rd IEEE International Conference, Boca Raton, 7–9 November 2011 : proceedings.* Boca Raton, Florida, USA, IEEE, 2011, pp. 393–398. DOI: 10.1016/j.asoc.2016.05.008
12. Arora M., Verma S. , Kavita An efficient effort and cost estimation framework for Scrum Based Projects, *International Journal of Engineering & Technology,* 2018, Vol. 7, No. 4.12, P. 52–57. DOI: 10.14419/ijet.v7i4.12.20992.
13. Business Process Model and Notation [Electronic resource]. Access mode: http://www.bpmn.org/
14. Estimate a Story [Electronic resource]. Access mode: https://www.quickscrum.com/ScrumGuide/175/sg-Estimate-A-Story
15. Saaty T. L. The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. London, McGraw-Hill, 1980, 287 p.
16. Crawford G., Williams C. The Analysis of Subjective Judgment Matrices : tech. report : R-2572-1-AF, Rand Corp. Santa Monica, Calif, 1985. [Electronic resource]. Access mode: https://www.rand.org/pubs/reports/R2572-1.html
17. Eduardo M. Sizing User Stories Using Paired Comparisons. [Electronic resource]. Access mode: https://www.researchgate.net/publication/222140271_Sizing_user_stories_using_paired_comparisons
18. Mel'nik K. V. Otsenka effektivnosti meditsinskoy informatsionnoy tekhnologii, *Matematicheskoye modelirovaniye protsessov v ekonomike i upravlenii innovatsionnymi proyektami (MMP-2013) : 1 Mezhdunarodnaya nauchno-prakticheskaya konferentsiya, Alushta, 9–15 Sentyabrya 2013 g. : tezisy dokladov*. Khar'kov, KHNURE, 2013, pp. 122–123.
19. Ovezgel'dyyev A. O., Petrov E. G., Petrov K. E. Sintez i identifikatsiya modeley mnogofaktornogo otsenivaniya i optimizatsii. Kiev, «Naukova dumka», 2002, 163 p.
20. Berteig M. Pitfall of Scrum: Excessive Preparation/Planning [Electronic resource], Access mode: http://www.agileadvice.com/tag/analysis-paralysis/
21. Lewis J. P. Fundamentals of Project Management. Warszawa, AMACOM, 2002, 128 p.

Received 15.01.2020.
Accepted 25.02.2020.

УДК 004.02

ТЕХНОЛОГІЯ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ПРИ ПЛАНУВАННІ СПРИНТУ

Мельник К. В. – канд. техн. наук, доцент кафедри Програмної інженерії та інформаційних технологій управління, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна.
Глушко В. Н. – Senior Consultant, Solution Architect, GlobalLogic, Харків, Україна.
Борисова Н. В. – канд. техн. наук, доцент кафедри Інтелектуальних комп'ютерних систем, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна.

АНОТАЦІЯ

Актуальність. У статті описується актуальний процес планування розробки програмного забезпечення, проблем планування і різні рішення цих проблем на основі використання методології Scrum.

Мета. Метою даної роботи є розробка технології для вирішення задачі планування спринту в умовах невизначеності і можливих ризиків з точки зору розробки програмного забезпечення.

Метод. Описано найбільш використовувані моделі життєвого циклу програмного забезпечення. Обґрунтовано вибір Scrum, як найбільш часто використовуваного представника гнучкою методології розробки програмного забезпечення. Проведено аналітичний огляд методів оцінки складності для історій користувача. Виділено основні проблеми планування спринту. Розроблено модель бізнес-процесу для реалізації IT-проекту по Scrum у вигляді BPMN-діаграми. Розроблено алгоритм вирішення проблеми планування Sprint Backlog в умовах невизначеності. Розглядається загальний процес вибору для користувача історій з Product Backlog для Sprint Backlog і шляхи вирішення можливих проблем. Формалізовані задача оцінки трудомісткості історій користувача і задача оцінки ризиків при плануванні. Була розроблена технологія вибору історій користувача для Sprint Backlog. Проведено чисельні дослідження технології підтримки прийняття рішень, яка була запропонована

DOI 10.15588/1607-3274-2020-1-14

в статті. Це дозволяє пропонувати її в якості практичного інструменту при плануванні спринту. Запропоновано метод оцінки адекватності запропонованої технології. Обрано набір ключових показників ефективності для оцінки продуктивності команди.

Результати. Була розроблена технологія планування спринту, яку можуть використовувати керівники проектів, власники продуктів і команди розробників для підвищення ефективності процесу прийняття рішень.

Висновки. Проведені експерименти підтвердили значимість запропонованої технології підтримки прийняття рішень і позволя рекомендувати її для практичного використання при плануванні програмних проектів. Наукова новизна полягає в поліпшенні процесу планування спринту за допомогою запропонованої технології, яка усуває невизначеність при визначенні трудомісткості користувальницьких історій і скорочує час, що витрачається на прийняття рішень.

КЛЮЧОВІ СЛОВА: спринт беклог, задача планування, невизначеність, трудомісткість історія користувача, задача вибору, ефективність команди.

УДК 004.02

ТЕХНОЛОГИЯ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПРИ ПЛАНИРОВАНИИ СПРИНТА

Мельник К. В. – канд. техн. наук, доцент кафедры Программной инженерии и информационных технологий управления, Национальный технический университет «Харьковский политехнический институт», Харьков, Украина.

Глушко В. Н. – Senior Consultant, Solution architect, GlobalLogic, Харьков, Украина.

Борисова Н. В. – канд. техн. наук, доцент кафедры интеллектуальных компьютерных систем, Национальный технический университет «Харьковский политехнический институт», Харьков, Украина.

АННОТАЦИЯ

Актуальность. В статье описывается актуальный процесс планирования разработки программного обеспечения, проблем планирования и различные решения этих проблем на основе использования методологии Scrum.

Цель. Целью данной работы является разработка технологии для решения задачи планирования спринта в условиях неопределенности и возможных рисков с точки зрения разработки программного обеспечения.

Метод. Описаны наиболее используемые модели жизненного цикла программного обеспечения. Обоснован выбор Scrum, как наиболее часто используемого представителя гибкой методологии разработки программного обеспечения. Проведен аналитический обзор методов оценки сложности пользовательских историй. Выделены основные проблемы планирования спринта. Разработана модель бизнес-процесса для реализации IT-проекта по Scrum в виде BPMN-диаграммы. Разработан алгоритм решения проблемы планирования Sprint Backlog в условиях неопределенности. Рассматривается общий процесс выбора пользовательских историй из Product Backlog для Sprint Backlog и пути решения возможных проблем. Формализованы задача оценки трудоемкости пользовательских историй и задача оценки рисков при планировании. Была разработана технология выбора пользовательских историй для Sprint Backlog. Проведены численные исследования технологии поддержки принятия решений, которая была предложена в статье. Это позволяет предлагать ее в качестве практического инструмента при планировании спринта. Предложен метод оценки адекватности предлагаемой технологии. Выбран набор ключевых показателей эффективности для оценки продуктивности команды.

Результаты. Была разработана технология планирования спринта, которую могут использовать руководители проектов, владельцы продуктов и команды разработчиков для повышения эффективности процесса принятия решений.

Выводы. Проведенные эксперименты подтвердили значимость предложенной технологии поддержки принятия решений и позволяют рекомендовать ее для практического использования при планировании программных проектов. Научная новизна заключается в улучшении процесса планирования спринта с помощью предлагаемой технологии, которая устраняет неопределенность при определении трудоемкости пользовательских историй и сокращает время, затрачиваемое на принятие решений.

КЛЮЧЕВЫЕ СЛОВА: спринт беклог, задача планирования, неопределенность, трудоемкость истории пользователя, задача выбора, эффективность команды.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Частка IT-індустрії в економіці України [Electronic resource]. Access mode: https://ua.112.ua/suspilstvo/chastka-it-industrii-v-ekonomitsi-ukrainy-stanovyt-4-vvp-kubiv-480452.html
2. Software Engineering Body of Knowledge (SWEBOK) [Electronic resource]. – Access mode: https://www.computer.org/education/bodies-of-knowledge/software-engineering
3. A Guide to the Scrum Body of knowledge (SBOKTM Guide) [Electronic resource]. – Access mode: http://www.scrumstudy.com/SBOK/SCRUMstudy-SBOK-Guide-2016.pdf
4. Sutherland J. Scrum: A revolutionary approach to building teams, beating deadlines and boosting productivity / J. Sutherland, J. J. Sutherland. – United States : Random House, 2014. – 256 p.
5. Agile Estimation Techniques – beyond Planning Poker [Electronic resource]. – Access mode: https://technology.amis.nl/2016/03/23/8-agile-estimation-techniques-beyond-planning-poker/
6. Agile Estimation Techniques: A True Estimation in an Agile Project [Electronic resource]. – Access mode: https://www.softwaretestinghelp.com/agile-estimation-techniques/
7. User stories complexity estimation using Bayesian networks for inexperienced developers / [J. Lopez-Martinez, A. Ramirez-Noriega, R. Juarez-Ramirez et al.] // Cluster Computing. – 2018. – Vol. 21. – P.715-728. DOI: 10.1007/s10586-017-0996-z
8. Karna H. Estimating software development effort using Bayesian networks / H. Karna, S. Gotovac // Telecommunications and Computer Networks (SoftCOM) : 23rd International Conference on Software, Split, 16–18 September

DOI 10.15588/1607-3274-2020-1-14

2015 : proceedings. Split: IEEE, 2015. – pp. 229–233. DOI: 10.1109/SOFTCOM.2015.7314091

9. Agile Evaluation of the Complexity of User Stories Using the Bloom's Taxonomy / [F. E. Barrera, M. A. García, H. G. González et al.] // Computational Science and Computational Intelligence (CSCI'17) : 4th Annual International Conference, Las Vegas, 14–16 December 2017 : proceedings. Las Vegas, Nevada: Conference Publishing Services (CPS), 2017. – P. 1047–1050. DOI:10.1109/csci.2017.182
10. Adapting Bloom's Taxonomy for an Agile Classification of the Complexity of the User Stories in SCRUM / [F. E. Castillo-Barrera, M. Amador-García, H. G. Pérez-González et al.] // Software Engineering Research and Innovation (CONISOFT'18) : 6th International Conference, San Luis Potosi, 24–26 October 2018 : proceedings. San Luis Potosi : IEEE, 2018. – P. 139–145. DOI: 10.1109/CONISOFT.2018.8645899
11. Nassif A. B. Estimating software effort based on use case point model using Sugeno Fuzzy inference system / A. B. Nassif, L. F. Capretz, D. Ho // International Conference on Tools with Artificial Intelligence (ICTAI 2011) : 23rd IEEE International Conference, Boca Raton, 7–9 November 2011 : proceedings. Boca Raton, Florida, USA: IEEE, 2011. – P. 393–398. DOI: 10.1016/j.asoc.2016.05.008
12. Arora M. An efficient effort and cost estimation framework for Scrum Based Projects / M. Arora, S. Verma, Kavita // International Journal of Engineering & Technology. – 2018. – Vol. 7, № 4.12. – P. 52–57. DOI: 10.14419/ijet.v7i4.12.20992.
13. Business Process Model and Notation [Electronic resource]. – Access mode: http://www.bpmn.org/
14. Estimate a Story [Electronic resource]. – Access mode: https://www.quickscrum.com/ScrumGuide/175/sg-Estimate-A-Story
15. Saaty T. L. The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation / T. L. Saaty. – London: McGraw-Hill, 1980. – 287 p.
16. The Analysis of Subjective Judgment Matrices : tech. report : R-2572-1-AF / G. Crawford, C. Williams / Rand Corp. – Santa Monica, Calif, 1985. [Electronic resource]. – Access mode: https://www.rand.org/pubs/reports/R2572-1.html
17. Eduardo M. Sizing User Stories Using Paired Comparisons. [Electronic resource] / M. Eduardo. – Access mode: https://www.researchgate.net/publication/222140271_Sizing_user_stories_using_paired_comparisons
18. Мельник К. В. Оценка эффективности медицинской информационной технологии / К. В. Мельник // Математическое моделирование процессов в экономике и управлении инновационными проектами (ММП-2013) : 1 Международная научно-практическая конференция, Алушта, 9-15 Сентябрь 2013 г. : тезисы докладов. – Харьков : ХНУРЭ, 2013. – С. 122–123.
19. Овезгельдыев А. О. Синтез и идентификация моделей многофакторного оценивания и оптимизации / А. О. Овезгельдыев, Э. Г. Петров, К. Э. Петров. – К. : «Наукова думка», 2002. – 163 с.
20. Berteig M. Pitfall of Scrum: Excessive Preparation/Planning [Electronic resource] / M. Berteig. – Access mode: http://www.agileadvice.com/tag/analysis-paralysis/
21. Lewis J. P. Fundamentals of Project Management / J. P. Lewis. – Warszawa: AMACOM, 2002. – 128 p.
22. Berteig M. Pitfall of Scrum: Excessive Preparation / Planning [Electronic resource] / Access mode: http://www.agileadvice.com/tag/analysis-paralysis/ , 17.01.2020.
23. James P. Lewis. Fundamentals of Project Management / P. Lewis James. – AMACOM, 2002. – 128 p. ISBN 978-0814471326.

DOI 10.15588/1607-3274-2020-1-14

UDC 004.942:316.454.54

DIAGNOSTIC SYSTEM OF PERCEPTION OF NAVIGATION DANGER WHEN IMPLEMENTATION COMPLICATED MANEUVERS

Nosov P. S. – PhD, Associate Professor of Navigation and Electronic Navigation Systems Department, Kherson State Maritime Academy, Ukraine.

Zinchenko S. M. – PhD, Senior Lecturer of Ship Handling Department, Head of the Laboratory of Electronic Simulators, Kherson State Maritime Academy, Ukraine.

Popovych I. S. – Dr. Sc., Full Professor of the Department of General and Social Psychology, Kherson State University, Ukraine.

Ben A. P. – PhD, Associate Professor, Deputy Rector for Scientific and Pedagogical Work, Kherson State Maritime Academy, Ukraine.

Nahrybelnyi Y. A. – PhD, Dean of the Department of Navigation at the Kherson State Maritime Academy, Ukraine.

Mateichuk V. M. – Assistant of Ship Handling Department, Head of the Laboratory of Electronic Simulators, Kherson State Maritime Academy, Ukraine.

ABSTRACT

Context. The article focuses on the question of automated decision-making analysis made by the operator in ergatic systems of critical infrastructures on the example of marine transport control in difficult navigation conditions. It is evident enough that the main criterion for an adequate perception of input information done by an operator is highly likely to predict the choice of behavioral decision-making strategies in discrete time conditions. However, the difficulty of modeling the operator's actions is found to be lying in non-linear pattern of taking definite decisions in emergency situations and deviations from the Codes and Rules.

Objective. The research purpose strategy of conducted investigation can be defined as the development of the mathematical platform for a decision support system (DSS) module with an aim to identify the class-forming set of atomic elements. In particular this issue determines the fact of distortion of the perception of information about navigation risks predicting the operator's behavior pattern while having vessel running process. This is possible to have it depicted through formal analysis.

Method. To capture the analysis of danger perception by the operator the paper introduces a mathematical model of data collection which identifies the fact of perception distortion in the form of attribute space of metadata obtained by the method of converting information from navigation devices. Besides, the factor of disorientation of the operator can be considered to be a shift on a displaced bridge which significantly affects on the analysis of information for adequate decision making. In addition, taking into account the failure of navigation equipment such as: RADAR, ARPA, AIS, ECDIS, especially while doing exit from the automatic control mode, a dangerous precedent can possibly be created for the operator not ready to perceive the complexity of the situation. To make it work a formal analysis was carried out using the extending risks possibility level tasks during the transition under these conditions. In addition to this item, a probabilistic model of perceiving the situation under the conditions of the error set is reported to have been constructed. So, as the result, the modeling process turned out to show the definite evidence of getting no way possibility to have the degree of criticality of the navigation situation determined without a clear identification of factors affecting the distortion of perception of the operator. Nevertheless, generalized statistical data are sure to be not enough and there is a special need of taking into account an individual information model of each operator for the effective work of DSS as this process faces real challenges. It must be significantly noticed that in order to analyze the perception of information by the operator a special test for defining preferences when choosing a strategy of control actions in the form of maneuvering under difficult navigation conditions purpose was created. Regarding the test results, as well as data on the passage of locations, certain attention is advised to be drawn to the classification analysis of 15 parameters using artificial neural networks having been carried out by our team and, as a consequence, the boundaries of deviations in the perception of navigational danger were found out and clarified. Additional superior item to be spoken about is certainly the introduction of rules and algorithms having been welcomed into the DSS core including the following: interaction field, RADAR and NIS synchronization tools; actual navigational hazard in a given cartographic area; ships trajectories and, as a result, simulations of probable deviations in the information perception of the operator.

Results. In order to meet beneficial agreement between the effectiveness of the developed DSS with the proposed formal-analytical approaches an experiment was assumed to be appropriate to be conducted using the Navi Trainer 5000 navigation simulator (NTPRO 5000). Based on the foregoing, due to comprehensive results in experiment metadata for the 2.5 years of operation of navigation simulators and DSS software tools the identification of the deviation probabilities in the information perception of dangers was achieved and export the predicted data to new locations for the operator and cartographic areas was performed. Undoubtedly, the experimental investigation confirmed the hypothesis of the study and reflected completely the feasibility of using this DSS to make predictions of possible risks when control the vessel by analyzing the information model of the operator.

Conclusions. Formal-analytical approaches presented in the study combined with the developed DSS software tools and the information itself made it possible to classify the decision-making strategies of the operator when control the vessel and to predict the probability of catastrophic consequences. The feasibility of the proposed models and methods was successfully revealed by carried out experiments.

KEYWORDS: decision support systems; operator information model; computer navigation simulators; probability of risks; human factor information analysis, automated control systems, automatic control systems.

DOI 10.15588/1607-3274-2020-1-15

ABBREVIATIONS

OOW is an officer on watch;

DSS is a decision support system;

NT 5000 is a navigation simulator "Navi Trainer 5000";

COLREG is an International Rules of Preventing Collision at Sea;

STCW is a Standards of Training, Certification and Watchkeeping;

ARPA is an automatic radar plotting aid;

AIS is an Automatic Identification System;

ECDIS is an Electronic Chart Display and Information System;

RADAR is a radar station;

CPA is a Closest Point of Approach;

TCPA is a Time to Closest Point of Approach;

IMO is an International Maritime Organization;

SOLAS is an International Convention for the Safety of Life at Sea;

ICAR 72 is an International Collision Avoidance Rules;

GPS is a Global Positioning System;

RBF is a Radial basis function;

LAT is a Latitude;

LON is a Longitud;

COG is a Course over ground;

SOG is a Speed over ground;

HDG is a Heading;

LOG is a Speed through water;

SET is a Drift direction;

DRIFT is a Drift;

SPD F is a Speed forward;

SPD A is a Speed Aft;

RUD is a Rudder angle;

ROT is a Rate of turn;

RPM L is a Revolutions per minute.

NOMENCLATURE

f is a navigation equipment information;

A is an ahead;

ξ is a rule ID;

U is a navigation tool;

M' is a navigational danger;

L is a turn on the left;

B is a backwards;

R is a turn on the right;

g_{ij} is a vessel position;

α is a strategy of movement;

$P_{ij}^{'}(\alpha)$ is a probabilities of movement;

y^* is a distance from the actual position;

y is a radius of the probable interaction;

$P_{ij}(x)$ is a probabilities of the adjusted direction;

φ is a mapping describes the displacement of the vessel;

G is a interaction radius;

β_{ij} is a framework of the distribution law;

μ_i is a maneuvering strategy;

μ_1 is a ship propeller control, smooth turning of wide radius;

μ_2 is a turning the steering wheel, smooth turning of the middle radius;

μ_3 is a use of bow thrusters, smooth turning of small radius;

μ_4 is a turn at anchor, sharp turn at ultra-small radius;

Y_i is a series of relations with respect to the operator;

Y_1 is a accident of possible collision of ships;

Y_2 is a accident of bulk near the berths of the strait;

Y_3 is a accident of ship grounding;

λ is a level of perception of navigation hazard;

R^* is a converter;

$\widetilde{*}$ is a displaced navigation bridge;

w is a distortion due to incomplete navigation watch;

S is a navigation situation;

Ω is a probable error in perceiving situation;

$\mathbf{G}$ is a vessel position matrix;

$\mathbf{W}$ is a matrix distortion due to incomplete navigation watch;

O is a navigation obstacle;

D^t is a set of the direction;

x^t is a given direction;

Q_i is an additive convolutio;

ω_i is a weights for maneuvering criteria.

INTRODUCTION

There is a widely-spread tendency in modern shipping practice for having minimization of sailing costs process accompanied with constant reduction of crews of vessels including navigational staff. Therefore, navigation and control of the ship are conventionally prioritized to be carried out in the context of reducing the number of people, as well as the lack of staff on the navigating bridge. Unfortunately, modern fleet is noticed to reflect the world experience of having nearly two or three people alone on the bridge in heavy navigation cases. In addition, the actual navigation and management of the vessel is mostly performed by one person, to be exactly, by the captain or officer of the watch. It must be taken into account that such a minimum number of people is reviewed to be the most profitable and advantageous way in having poor navigation. Moreover, difficult conditions of emergency situations constantly require additional number of crew members on the last boarder of safe level.

So, to be more precise, general formula of number of peoples to be operating on the bridge can possibly look as follows: OOW; OOW + 1 or Master + OOW + 1.to be.

It must be significantly noticed that in the conditions of entry/exit of the vessel in/from the port, a pilot is sure to be added (if local laws do not permit another scheme). Special emphasis to be done is that the pilot from one point of view is not a responsible person and, from another one, according to all international standards, really

DOI 10.15588/1607-3274-2020-1-15

is a situational member of the navigational team on the bridge. Being ultimate adviser and assistant to the captain he gets used to having additional skills and knowledge in a particular water area. In spite of all mentioned above, the situation seems to be following a pattern according to which most pilots do not perceive themselves to be assistants in this way considering mostly themselves to be an independent leader on the navigation bridge. Such poorly done actions could possibly cause tension between the pilot and the ship's captain and could create negative air on the bridge leading to unpredicted consequences, emergencies and, as a result, unwelcomed stress [1].

It can be clearly seen, speaking about the prioritizing task of vessel moving in the port areas, that each ship is said to have a motion vector in one of three directions directly on the course or in reverse. In spite of this, port areas, passing locations and narrows try to avoid usual practice of the vessel movement in the preferred direction. As the presence of insurmountable cartographic obstacles on the way constraints to the draft of the vessel or a significant amount of maritime traffic there is definitely observed fundamental approach of the operator to propose the altering direction of movement choosing the one on which the obstacles are considered to be minimal [2].

However, being involved into choosing a maneuvering strategy the operator processes with the parameters of movement and location of the vessel based on his own experience [3]. The situation is recognised to be especially disadvantageous becoming aggravated at the time of completion of the dynamic positioning modes and inappropriate watch keeping. These issues leads to the worsening of the process of adequate perception of navigation dangers. It must be emphasised that abruptness and precipitousness of the situations are contributing to the distortion in the risks assessment during the adoption of complex decisions being limited by the time of maneuvering the vessel. Summing up all spoken about ,it might be noted that there is definitely a tendency of decreasing level of the safety control in maritime transport to be observed investing to the increases of the probability of shipwrecks.

The emphasis must be placed to the fact of having real troubles in decreasing the occurrence of violations experience leading to distortions in the perception of danger by operators. This item is extremely difficult to be determined without a psychologist involvement into the participation which is a problematic one because of his not being a part of the ship's crew moment. This issue could be possibly solved by the introduction of a specially developed information system which provides control and prevention of negative consequences services.

The object of research is said to be the process of automated identification of distortions in the perception of risks.

The subject of research are models and algorithms implementing the process of automated identification of distortions in the perception of navigational hazards.

The aim of the study is to develop the mathematical platform for a decision support system (DSS) module to identify the class-forming set of atomic elements that determine the fact of distortion of the perception of information about navigation risks through a formal analysis and prediction of the operator's behavior when control a vessel.

The purpose of article is said to be solving the following tasks:

1. To analyze the probabilistic models of the transition of the vessel on the cartography of the location with limitations of the navigation risks relatively to *f* and the international rules for the management of the vessel according to *A* when have a watch. To determine the interrelation of ξ and *v* relatively to the bias *U*, which made the DSS signal, *R* ξ possible to be revealed.

2. To define in a formal form an incomplete model of perception by operator of M' which is characterized by inaccuracy and dependence on variables. This item is reported to require the development of the identification mechanism by means of DSS.

3. To develop a local metric for the transition of the vessel to a new state based on a set of directions: *A*, *R*, *C*, *L*, *B* – defined by the boundaries $g_{i+1,j}$ and probabilities of movement $P'_{ij}(\alpha)$ depending on *y**, which will determine the probability of the moving vessel in $P_{ij}(x)$.

4. To determine the DSS operation scheme that will provide possibility at a certain given time interval to define discretely the a priori probabilities of the ship's movement $\varphi = \varphi_2 \circ \varphi_1$ in the nearest *G*, proposing several samples of identification of the probable interaction of the ship with navigational hazards β_{ij} at the current time at $g_{ij} = 1$.

5. To develop an adaptive testing algorithm for cadets aimed to identify operator preferences in the form of μ_i unveiling influence on deviations in the perception of Y_i, based on Pareto principles of optimal alternatives.

6. To have an experimental investigation done by automated analysis of logfile data of perform maneuvers in the narrowness of the Bosphorus Strait using artificial neural networks.

7. To determine metadata for DSS with an attempt to create an individual operator model designing to be in the form of a preference map based on the analysis of server data of the NTPRO 5000 navigation simulator.

Thus, to be precise, the review papers domain goal with the implementation of all aforesaid tasks can be defined as identification of the stable models of operator behavior developed in experience in certain situations, relative to locations, COLREG rules and ship maneuverability. This detailed information would definitely enable DSS to propose key contributions into the enhancement of the accuracy level of forecasting critical situations.

1 PROBLEM STATEMENT

Significant new insights to the definition of the probability of shipwrecks in case of bridge watchkeeping incomplete configuration must be offered and as a result,

DOI 10.15588/1607-3274-2020-1-15

the information tools to prevent them might be practically determined. To have several ambiguities avoided in such situations the rules of the COLREG and STCW are sure to be carefully followed when carrying out the navigation watch.

It must be emphasised that for having these purposes achieved an information module was developed in the previous research papers determining the composition of the watchkeeping team in real time [4] (Fig. 1 a, b). In addition to the spoken above, heart rate and temperature sensors were made use of, indicating the probable stressful condition of the chief officer and second officer on the bridge (Fig. 1 b).

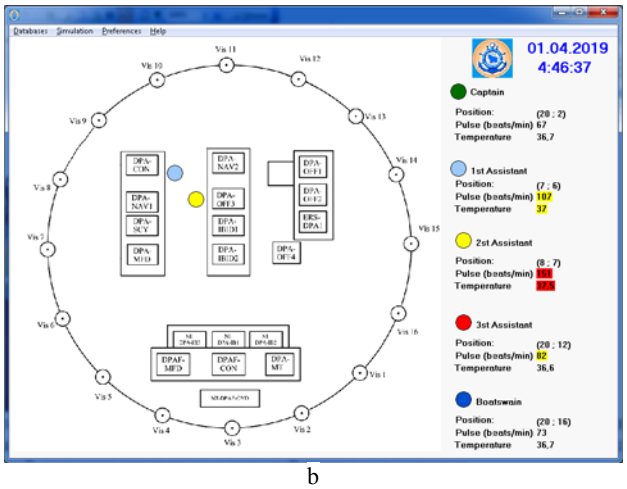
Besides, one more highlighted subject to be paid attention to is tracking the position of crew members due to its being a significant use of while defining the interaction models within the team. Notwithstanding this item, it is troubled enough to determine the forms and boundaries of the perception of the navigation situation by the operator who is responsible for decision making.

The primary perception model of the navigation situation: $\lambda = Sf + w$.

The main problematic action is the necessity of additional information tools to meet the needs of the reliability in identification of the navigation situation.

For example, let us take the Radar Station as a device *U*, whose output signal gives the values of the parameters of the navigation situation under study.

a



b

Figure 1 – Incomplete watchkeeping team on the bridge

Then the task of interpreting measurement (1) is reduced to a transformation λ to the form characteristic of measurement on the instrument *U*. At the same time, the designed DSS should have such a converter *R**, which allowed generating the signal $R^*\lambda$, as the most accurate *Uf*.

Summing up, basing on this model of perception of the navigation we could definitely come up to the conclusion that using a number of synchronized navigation devices such as Radar Station, ARPA, AIS, ECDIS and lack of deviating from the rules of watch keeping lead to beneficially leveled down navigation risks.

However, taking into account the case of poorly performed navigation watch, especially being complicated by the fact of the $\tilde{*}$, it is necessary to build a model under conditions of probable error in Ω and, as a result, the risks of catastrophic consequences would seem to be increasing (Fig. 2 bridges).

Contemporaneously, the control of navigation equipment cannot be completely rely on as the time ranges of loss of system interaction with the instruments may drive to untimely decisions-making strategy when controlling the vessel.

Consider this situation in the mathematical aspect [5, 6].

Let us take an example when the operator analyzed the navigation situation under conditions of incomplete watch keeping and its perception is $\tilde{\lambda} = \tilde{\Omega} f + \tilde{w}$, at the same time, this incomplete model of perception $\tilde{\lambda}$, deliberately bears an inaccuracy, then: $\tilde{\lambda} = Sf + w \in \Re_n$ and $\tilde{\lambda} = \tilde{\Omega} f + \tilde{w} \in \tilde{\Re}_{\tilde{n}}$, where $f \in \Re_m$. So, an incomplete perception model is clearly seen to be followed such as: $M' = \left[\begin{pmatrix} S \\ \tilde{S} \end{pmatrix}, \begin{pmatrix} \Theta & 0 \\ 0 & \Theta \end{pmatrix}\right]$, where $\tilde{S} \neq \tilde{\Omega}$, and it's statistically correct to deduce that the distribution is: $\begin{pmatrix} w \\ \tilde{w} \end{pmatrix} \to \aleph\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Theta & 0 \\ 0 & \Theta \end{pmatrix}\right)$.

Apparently, it is to be emphasized that without having clear identification of the variables *w* and $\tilde{w}$, the process of determination of the degree of perception of the criticality of the navigation situation is looked at as being not a successful one. Therefore, the generalized statistical data as on the individual operator model can be regarded as having a problematic issue to have reliance in [7].

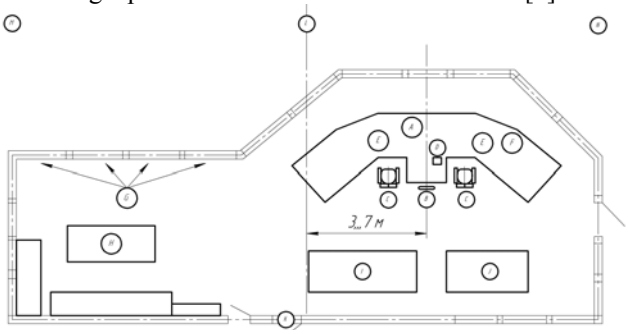


Figure 2 – Offset Bridge Layout

DOI 10.15588/1607-3274-2020-1-15

The complexity of making such a model work under the conditions of a number of difficult – to be -determined factors introduces insurmountable obstacles to resolve this problem. In view of above, the development of a decision support system is to obtain an essential value of with an aim of constant monitoring of situations and risks associated with them while keeping navigation watch on the bridge.

It is to be highlighted that operators of maritime transport are in beneficial position due to the monitoring possibility being carried out through RADAR and Nautical Equipment [8] in a location at a *r*, and the opportunity to choose the direction in which they observe the least amount of maritime traffic not speaking about the absence of cartographic obstacles.

In the circumstances of drawing an increased attention to, the operator constantly monitors the field of interaction, which is a combination of two matrices (**G***;* **W**). At the same time g_{ij}, the value corresponding to the presence or absence of the vessel in this position: $o_{ij} = \{0;1\}$ the value corresponding to the presence (1) or absence (0) of navigation hazard in this cartographic region.

We introduce a metric on cartography in the form of frames of the nearest radius of movement (Fig. 3.), the direction of motion represents: $D^t = \{A, B, L, R\}$.

The operation of moving the vessel is denoted by a variable α taking the values *A, R, L, B*, while the boundaries of the frames in the radius of movement will assume that: $g_{i+1,\,j} = g_{ij}^{(A)}$, $g_{i,\,j+1} = g_{i\,j}^{(L)} \ldots\ g_{i\,j} = g_{i\,j}^{(C)}$.

For frames with $g_{ij} = 1$, we define the probabilities of movement, provided that the vessel moves to the open frame: $P_{i\,j}^{'}(\alpha) = \left(\left(1 - g_{i\,j}^{(\alpha)}\right)\left(1 - o_{i\,j}^{(\alpha)}\right)\right)/4$.

Besides, by means of automatic recognition of dangerous targets of the Radar Station [9] the function of analyzing the surrounding navigation environment has been introduced.

For example, let us determine the a priori probabilities of the ship moving in the nearest radius of interaction. Simultaneously to this, the cartographic obstacles are taken to be expressed as constants and the moving targets could be distributed depending on the perceived navigation hazard:

$$P_{i\,j}^{''}\left(D^t\right) = 1 - \left(\left(\sum_{k=1}^{y^*} g_{i,\,j+k} + y - y^*\right)\Big/ y\right) P_{i\,j}^{'}(A),$$

where y is the radius of the probable interaction of the vessel with navigational hazards.

y*, can be considered as a derivative of CPA and TCPA, and $P_{ij}^{'}(\alpha)$. The next step to be done is the increase of probability of moving in a given direction $x^t = \{x^{'}, \ldots, x^{'''}\}$.

$$P_{ij}(x) = P_{ij}^{''}(x) + \alpha \min\left[1 - P_{ij}^{''}\left(x^t\right)\right],$$
$$P_{ij}\left(x^t\right) = P_{ij}^{''}\left(x^t\right) - \left(\alpha \min\left[1 - P_{ij}^{''}\left(x^t\right)\right]/4\right).$$

At the same time, $\alpha \in [0;1]$ might be defined as the strategy of movement (maneuvering, the chosen course of the vessel motion, speed).

However, despite of the foregoing, the idea of the metric of real maneuvers assuming not a transition from one frame to another (if this frame is empty)is sure to be accentuated. Moreover, the actions of the operator thus lead to the movement of the vessel in definite time being the choice of strategy. So, new contemporary approaches relating this subject are highly likely to be found out to make it possible to forecast the actions of the operator with sufficient accuracy and the probability of having warning of disasters just in time.

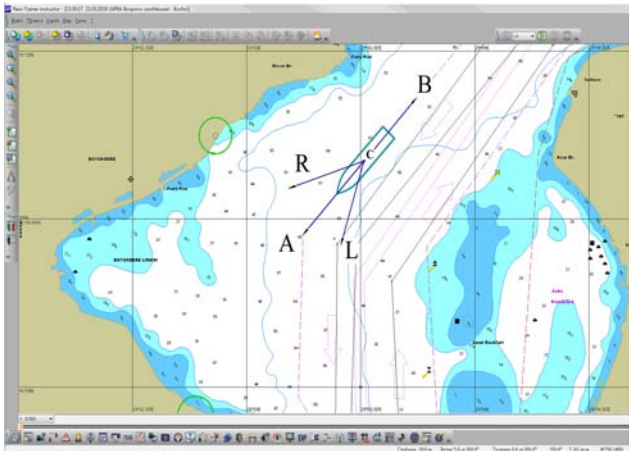


Figure 3 – Ship transition probabilities

2 LITERATURE REVIEW

The dominating strategy in the attempt to increase the efficiency of maritime transport is having capable navigational bridge team [10] taking into account the latest research on the mental state of expectations of professional activities [11]. This way is efficient enough due to safety arrangements in accordance with international standards and the regulations of IMO, SOLAS, STCW, COLREG [12]. As a rule, it is precisely defined that dominating aspect is the negative manifestation of the human factor that affects the outcome of complicated maneuvers when passing locations during the watchkeeping on the captain's bridge [13–16]. Besides, the factors directly affecting each operator [17–19] are obviously not the single ones to have essential impact on, there are factors of the organization of interaction between members of the watch [20–21]. In addition to cartographics and weather conditions [22], certain number of factors depending on insufficient qualifications of the crew becomes vitally important [23]. According to the carried out research, it is to be underlined that the quality of maneuvering is directly affected by the number of simultaneous information signals causing troubles connected with the perception threshold exceeding [16]. However, this investigation is facing

DOI 10.15588/1607-3274-2020-1-15

challenges while examining the other side of the situation, namely the distortion of the perception of the shipwreck hazard. These deviations from an adequate perception of navigational danger cause confusing actions to happen being significantly influential on the path of the vessel and, in sum, are likely to generate tragic consequences. Models and methods in the framework of set theory, theory of logic, operations research, game theory, and probability theory can happen to be of great assistance being an operator mathematical description of the decision making process in this area of research [24–26]. The reducing of the composition of the bridge team practically means the expansion risk of a catastrophic situation making the navigation process troublesome to be predicted. Therefore, it would be beneficial to assume that there is a certain need in deeper understanding of the way of the decision-making process of the operator which is greatly influenced by complex formalized factors impossible to be tracked without the help of a qualified psychologist. In addition, in case of incomplete configuration of bridge team, the operator is the only person to be involved into the decision making process and in case of an erroneous action none of the members of the watch will be able to prevent these events from happening [27]. However, there are no vivid evidences in having references to systems able to identify these factors in an automatic way in order to exclude the possibility of occurring disasters in the notorious literature.

3 MATERIALS AND METHODS

Based on the foregoing, the purpose of the DSS development requires to be abstract from movements relative to the course of the four options to a much larger area in terms of the time limits for the implementation of maneuvers.

Having number of possible actions expanded could be considered as a problematic issue causing taking into account only relatively typical maneuver .

Regarding that the classic options for avoidance of collision in maritime transport are rare, there may be quite a lot of possible actions. The actual question is about the choice of the most suitable one [28–29].

In fact, the main thing that constitutes a problem of the formal description presents the idea of each operator having his own experience in performing certain maneuvering operations. In this regard, a task comes up in setting up the experience connections that is to be strengthened by the practice of navigation relative to each operator.

The calculated probabilities for each direction of movement determine the state of the interaction model at each stage. The evolution of the G is represented by the recurrence relation: $G_{n+1}= \varphi(G_n)$, where $\varphi = \varphi 2 \circ \varphi 1$. Then for each interaction frame, such that $g_{ij} = 1$, we introduce a variable β_{ij} within the framework of the distribution law.

The mapping φ1 describes the displacement of the vessel to free cartographic areas within the framework of the COLREG rules. It does not obligatory mean the likelihood of becoming this area the target of maneuvers of other vessels. In turn, the mapping φ2 is introduced in order to define the course combinations of the vessels in the nearest range of actions. In order to resolve the effects of mappings φ1 and φ2 on the considered formal model, we present the following sequence of DSS work (Fig. 4).

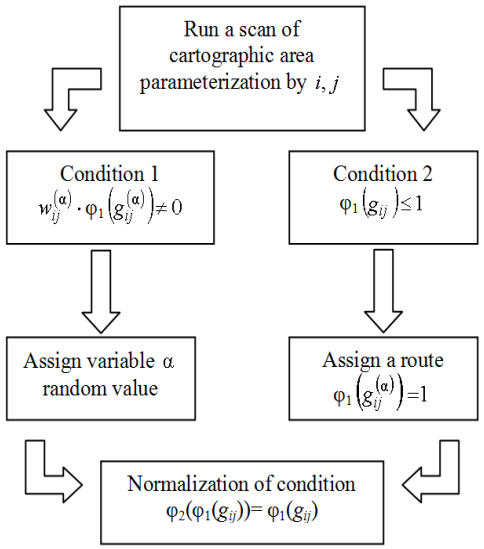


Figure 4 – DSS workflow for identifying probabilistic risks

However, the real ambiguity is that the danger lies within the possible operator fallaciousness who is likely to perceive the navigation situation in his own way and, as a result, the risk of a crash in case of a chosen trajectory is triggered to get the wrong turn. Consequently therefore, the main factor influencing on taking the wrong strategy direction is supposed to be stable experience connections. This issue possesses inconsistent nature when taking into account the bundle of existing rules and recommendations of international shipping regulations.

Hence, in such case, the task of the DSS that needs to be accomplished is to clearly identify the data of destructive relationships from the scratch while performing simulator training on cadets or undergoing re-certification by a crew having more than 10 years of experience.

For the particular purpose of this item, an algorithm was developed and taken to include a specialized adaptive test and synchronization tools with navigation simulators. Being of a particular importance, with regard to getting these issues accomplished, they are aimed at identifying the investigated deviation factors in the perception of navigation danger.

This algorithm provided the possibility to have been testing 351 cadets of the Kherson State Maritime Academy for more than two years of navigational tasks. The results depicted that at the time of making a decision from three to four possible maneuvering strategies of the vessel happen to be made mostly use of. It must be added that this set of strategies is due to positive experience and does not meet any contradictions with the rules of the ICAR

DOI 10.15588/1607-3274-2020-1-15

72. Moreover, there are a number of criteria regarding the performance functions of the maneuver to complete a complex turn using the example of the Bosphorus Strait. Thus, a set of alternatives for decision making is formed $Y_1,\ldots, Y_3$.

In a survey of cadets and experienced operators, three performance criteria being inverse to the risks of shipwrecks are named to be the most clearly distinguished.

So, we need a mathematical model to be constructed in terms of the Pareto efficiency theory [30].

It is worth mentioning that DSS during multiple passage of locations and performing maneuvers finds out the most widely spread behavior pattern of the cadet (operator) in typical situations. Apparently, it is true to say that multiple repetition of successful maneuvers in similar situations would lead to reinforcement of the situation-maneuver combination in reference to each operator. Thus, an array of preferences is formed regarding each alternative within the navigation situation.

So during the experiment, the DSS determined the following series of relations in reference to the operator:

$$Y_1:\ \mu_1 \approx \mu_2;\ \mu_2,\mu_1 \succ \mu_3;\ \mu_3,\mu_2,\mu_1 \succ \mu_4,$$

$$Y_2:\ \mu_2 \succ \mu_1;\mu_2 \approx \mu_3;\mu_4 \approx \mu_1;\mu_3 \succ \mu_1;\mu_2,\mu_3 \succ \mu_4,$$

$$Y_3:\ \mu_2 \succ \mu_1;\mu_2 \succ \mu_3;\mu_2 \succ \mu_4;\mu_3 \approx \mu_4;\mu_1 \succ \mu_3;\mu_1 \succ \mu_4.$$

As for the operator himself, he fronts the question of having the most suitable option to be chosen from taking into account the criteria, using Q_1, Q_2.

For Q_2, weights were selected for the criteria regarding the experience of performing maneuvers: $\omega_1 = 0{,}5;\ \omega_2 = 0{,}2;\ \omega_3 = 0{,}3$.

For each of the criteria, the DSS constructs a relationship matrix. $Y_{1,\ldots,3}$.

Introduce the rule:

$$\xi_Q(\mu_i,\mu_j) = \begin{cases} 1, & if\ \ \mu_i \succeq \mu_j\ \ \text{or}\ \ \mu_i \approx \mu_j \\ 0, & if\ \ \mu_i \prec \mu_j. \end{cases}$$

Then, the relationship matrix $Y_{1,\ldots,3}$ will have the form:

$$\xi_{q_i}(\mu_i,\mu_j)_{Y_1} = \begin{vmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{vmatrix}, \text{ for } \omega_1 = 0.5,$$

$$\xi_{q_i}(\mu_i,\mu_j)_{Y_2} = \begin{vmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{vmatrix}, \text{ for } \omega_1 = 0.2,$$

$$\xi_{q_i}(\mu_i,\mu_j)_{Y_3} = \begin{vmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{vmatrix}, \text{ for } \omega_1 = 0.3.$$

Convolution of relations will be equal:

$$Q_1 = Y_1 \bigcap Y_2 \bigcap Y_3 = \begin{vmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{vmatrix}.$$

Then the subset of non-dominated alternatives will be equal:

$$\xi'_{Q_i}(\mu_i) = 1 - \sup_{\mu_j \in \Sigma}\left\{\xi_{Q_i}(\mu_j,\mu_i) - \xi_{Q_i}(\mu_i,\mu_j)\right\};$$

$$\xi'_{Q_i}(\mu_1) = 1 - \sup\{1-0;0-0;0-1\} = 0,$$

$$\xi'_{Q_i}(\mu_2) = 1 - \sup\{0-1;0-1;0-1\} = 1,$$

$$\xi'_{Q_i}(\mu_3) = 1 - \sup\{0-0;1-0;0-1\} = 0,$$

$$\xi'_{Q_i}(\mu_4) = 1 - \sup\{1-0;1-0;1-0\} = 0.$$

In this way, $\xi'_{Q_i}(\mu_i) = [0;1;0;0]$.

Find Q_2:

$$\xi_{Q_2}(\mu_i,\mu_j) = \sum_{j=1}^{m} \omega_j\, \xi_j(\mu_i,\mu_j), \sum_{j=1}^{m} \omega_j = 1, \omega_j \geq 0.$$

The additive convolution of relations $Y_{1,\ldots,3}$ will be equal to:

$$\xi_{q_i}(\mu_i,\mu_j)_{Y_{1,\ldots,3}} = \begin{vmatrix} 1 & 0.6 & 0.6 & 1 \\ 0.4 & 1 & 0.7 & 0.4 \\ 0.4 & 0.7 & 1 & 0.4 \\ 0.6 & 0.6 & 0.6 & 1 \end{vmatrix}.$$

7. Find a subset of non-dominant alternatives for Q_2:

$$\xi''_{Q_2}(\mu_i) = 1 - \sup_{\mu_j \in \Sigma}\left\{\sum_{j=1}^{m} \xi_{Q_2}(\mu_j,\mu_i) - \xi_{Q_2}(\mu_i,\mu_j)\right\};$$

$$\xi''_{Q_2}(\mu_1) = 1 - \max\{(1-0.5);(0.2-0.7);(0.2-1)\} = 0.5,$$

$$\xi''_{Q_2}(\mu_2) = 1 - \max\{(0.5-1);(0.2-1)(0-1)\} = 1,$$

$$\xi''_{Q_2}(\mu_3) = 1 - \max\{(0.7-0.2);(1-0.2);(0.3-1)\} = 0.2,$$

$$\xi''_{Q_2}(\mu_4) = 1 - \max\{(1-0.2);(1-0);(1-0.3)\} = 0.$$

In this way, $\xi''_{Q_2}(\mu_i) = [0.5;1;0.2;0]$.

The last step to be spoken about in determining the choice of a maneuvering strategy by the operator is to find the intersection of the sets Q''_1, Q''_2 i.e., $Q'' = Q''_1 \bigcap Q''_2$.

As a result, we obtain a decision-making model for the operator regarding the four strategies considered:

$$\xi''_Q(\mu_i) = \min\left\{\xi''_{Q_1}(\mu_i), \xi''_{Q_2}(\mu_i)\right\}''_Q = \begin{vmatrix} \mu_1 = & 0.5 \\ \mu_2 = & 1 \\ \mu_3 = & 0.2 \\ \mu_4 = & 0 \end{vmatrix}.$$

DOI 10.15588/1607-3274-2020-1-15

t follows from the model that for this operator the strategy "μ_2 – turning the steering wheel feather, smooth turning of the average radius" will be most acceptable, strategies μ_1 and μ_3 are significantly less likely, and μ_4 is not considered at all acceptable.

Figure 5 – Increased perception difficulty due to worsening weather conditions

Then, the next stage of the DSS to be determined is the ratio specification of the expected actions of the skipper. Nevertheless, it could be a problematic one due to unpredictable nature of his decision-making strategy. To have the DSS operating on more profound level it is necessary to reach an agreement on specific terms of an experiment conduct with a sample sufficient for adequate conclusions. This effect may possibly be increased with changing weather conditions (Fig. 5). Such cognitive connections are likely to be traced from the very starting points of vocational training and certification. This particular practice having valuable and advantageous nature will definitely help to prevent negative consequences [31].

4 EXPERIMENTS

Essential stress must be added to the fact that to cover the aim of creating models of the DSS skipper's behavior, an analysis of the passage of the Bosphorus Strait with a view to completing a complex turn near Sariyer was sampled and successfully accomplished. Latitude and longitude data were processed with a discrete step of 5 seconds and GPS positioning – 1 meter (Table 1).

So, in order to obtain and effectively classify the trajectories by curvature statistical analysis of the data was under specific consideration to support the previous statement.

For sufficient accuracy, a second-level test site was applicable (Fig. 6).

Table 1 – Fragment of DSS database for the analysis of the maneuver trajectory

	TIME_	LAT	LON	COG	SOG	HDG	LOG	SET_
	505	40,93756362	28,96702804	192	20,194	192	20,194	282
	506	40,93664822	28,9667649	192	20,195	192	20,195	282
	507	40,93573194	28,96650331	192	20,197	192	20,197	282
	508	40,93481561	28,96624328	192	20,198	192	20,198	282
	509	40,93389934	28,96598482	192	20,172	192	20,172	282
▶	510	40,932983	28,96572793	192	20,201	192	20,201	281

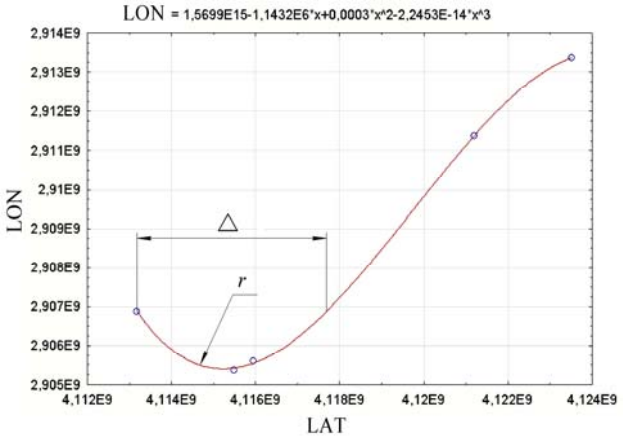


Figure 6 – Maneuver curvature graph

The resulting graph determines the delta dimension $\Delta = (4.113E9 \div 4.117E9)$, which allowed us to determine the curvature of the ship's trajectory.

The next step involved raising awareness of identifying four types of trajectories according to the chosen maneuvering strategies (Fig. 7–10).

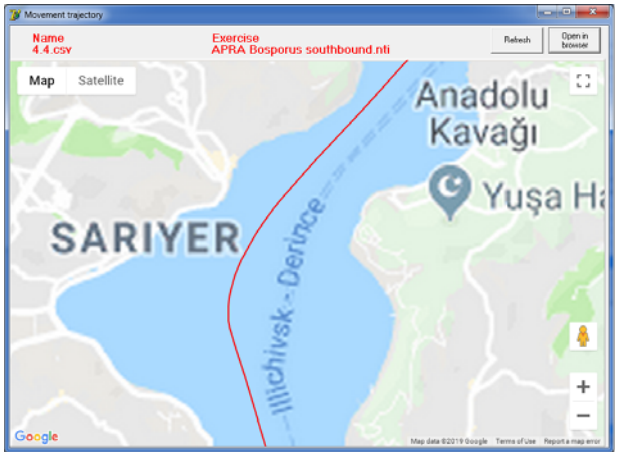


Figure 7 – Turning in a wide arc, μ_1

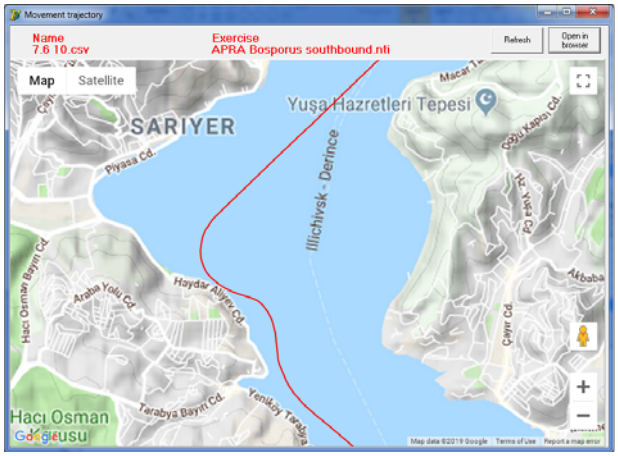


Figure 8 – Medium turning radius, μ_2

DOI 10.15588/1607-3274-2020-1-15

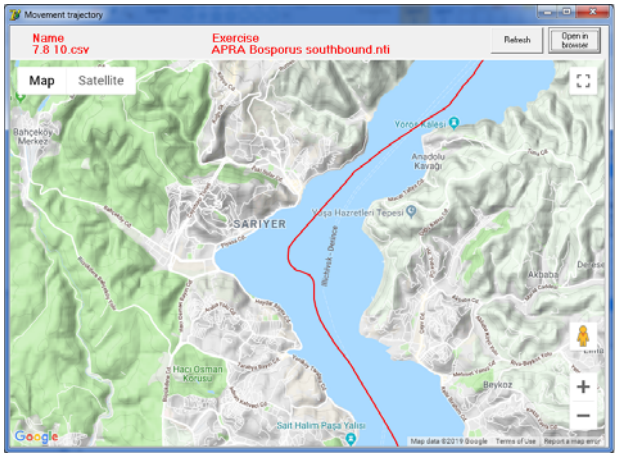


Figure 9 – Small radius turning, μ_3

Figure 10 – Turning with the anchor drop, μ_4

The initial analysis according to the scattering diagram showed the following results (Fig. 11).

Thus, it was established that it is necessary to take the data slice of the server of the NTPRO 5000 navigation simulator in latitude at 41.150000.

On this line, the turn maneuver is depicted to be performed directly; therefore, a series of data is the most relevant.

To be precise, the slice for each of the 34 experiments is declared to contain 15 parameters (Table 2).

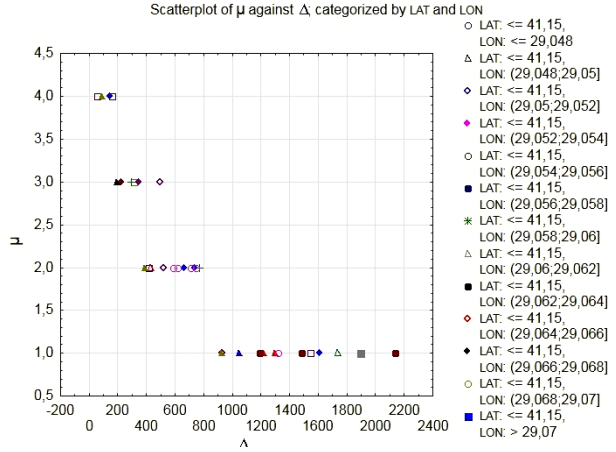


Figure 11 – Multiple scatter diagram of ship locations relative to strategies

According to the provided above table, as it can be seen, the graph indicates a significant spread in Δ and even the overlap between μ_1, μ_2 and μ_3. Therefore, other essential aspect for presenting accurate classification, as it deems necessary and advisable, is to get benefit of using deeper nonlinear classification methods.

Nevertheless, the initial stage with 2000 epochs and 5 hidden neurons did not seem to exhibit high efficiency (Table 3).

It is worth mentioning that being one of the most powerful mathematical apparatuses automated artificial neural networks proposes you to be allowed to work with multi-criteria and massive experimental data such as in this case [32–33].

Upon the conclusion of the first stage in the second one which is named as the use of artificial neural networks the parameters were significantly increased: the training covered fifteen networks (which is the maximum number for this experiment); the number of eras was 5000; the number of hidden neurons – 15; a radial basis function was chosen.

Table 3 – Results of the primary classification stage

Nets	Architecture	Performance	Performance Control	Performance Test
11	RBF 14-12-4	41,666667	50,000000	50,000000
12	RBF 14-12-4	66,666667	0,0000000	0,0000000
13	RBF 14-12-4	50,000000	50,000000	0,0000000
14	RBF 14-12-4	41,666667	50,000000	100,00000
15	RBF 14-12-4	75,000000	50,000000	50,000000

In this case, the results of the classification are noticed to show high efficiency level in training networks in the range 16.RBF (14154) – 18 RBF (14154) having the last one as the most effective among them (Table 4).

The confidence level was found out to be 0.999996 being able to be classified as a high indicator for the results (Fig. 12). To some extent, such an indicator can be considered to be a satisfactory one for the experiment and can be taken as a basis.

Table 4 – μ_i (Summary of classification) Educational

Architecture	μ –1	μ –2	μ –3	μ –4	μ –all
16.RBF 14-15-4	7,0	4,0	3,0	2,0	16,0
Right	7,0	4,0	2,0	1,0	14,0
Wrong	0,0	0,0	1,0	1,0	2,0
Right (%)	100,0	100,0	66,6	50,0	87,5
Wrong (%)	0,0	0,0	33,3	50,0	12,5
17. RBF 14-15-4	7,0	4,0	3,0	2,0	16,0
Right	7,0	3,0	3,0	2,0	15,0
Wrong	0,0	1,0	0,0	0,0	1,0
Right (%)	100,0	75,0	100,0	100,0	93,75
Wrong (%)	0,0	25,0	0,0	0,0	6,25
18.RBF 14-15-4	7,0	4,0	3,0	2,0	16,0
Right	7,0	4,0	3,0	2,0	16,0
Wrong	0,0	0,0	0,0	0,0	0,0
Right (%)	100,0	100,0	100,0	100,0	100,0
Wrong (%)	0,0	0,0	0,0	0,0	0,0

DOI 10.15588/1607-3274-2020-1-15

Table 2 – Data for analysis

LAT	LON	COG	SOG	HDG	LOG	SET	DRIFT	SPD F	SPD A	RUD	ROT	RPM L	Δ	μ_i
41.15007	29.0694	180	13.679	183	13.659	93	0.6	–0.278	–1.019	1	17	198	1543	1
41.15021	29.05515	191	17.849	188	17.812	277	1.1	–0.635	2.899	–15	–31	119	518	2
41.15027188	29.05443907	186	18.152	181	18.073	270	1.7	–0.389	3.749	–8	–36	120	931	1
41.1501253	29.05949366	163	19.565	161	19.557	251	0.6	0.041	1.1	0	–14	175	621	2
41.15003105	29.05948225	186	17.47	181	17.397	271	1.6	–0.155	3.31	0	–30	118	589	2
41.1502799	29.05739313	182	19.025	180	19.015	270	0.6	–0.466	1.749	1	–31	174	1216	1
41.15009557	29.06968935	187	13.3	187	13.29	110	0	–0.027	–0.058	1	0	196	162	4
41.15009869	29.05354389	144	14.475	143	14.472	233	0.3	0.584	–0.072	15	5	114	189	3
41.15016964	29.05770643	134	13.369	137	13.355	46	0.6	1.142	–2.407	7	49	174	151	4
41.15016259	29.05944346	163	15.011	164	15.011	73	0.1	–0.047	–0.225	–2	1	100	1322	1
41.15205802	29.05609934	176	15.444	178	15.44	87	0.4	0.678	–1.393	1	29	174	1045	1
41.15010354	29.05782304	178	11.161	179	11.16	88	0.2	–0.058	–0.262	0	1	99	741	2
41.15051669	29.05719364	207	13.925	195	13.61	284	2.9	–1.02	6.794	–35	–69	115	293	3
41.15037118	29.06389709	195	15.277	206	15.001	116	2.9	0.487	–6.228	5	94	174	226	3
41.15007018	29.05687433	178	16.997	174	16.968	264	1	–0.762	2.756	–10	–49	174	667	2
41.15007823	29.05126307	168	15.818	160	15.664	249	2.2	–0.369	4.725	–8	–45	117	428	2
41.15012971	29.05769572	171	10.845	171	10.844	81	0.1	0.159	–0.265	7	3	72	1615	1
41.15008146	29.06489771	180	9.098	179	9.094	266	0.1	–0.029	0.23	–1	–6	157	1732	1
41.15057277	29.06263983	167	18.085	167	18.084	257	0.1	–0.227	0.376	2	–8	174	89	4
41.15040316	29.04961019	174	9.064	157	8.686	247	2.6	0.168	5.013	–35	–38	90	414	2
41.15048491	29.06069001	180	14.069	183	14.06	92	0.6	–0.476	–0.732	–11	5	193	352	3
41.15056649	29.06322449	201	15.433	211	15.184	121	2.8	1.019	–6.545	18	106	174	926	1
41.15009548	29.05537871	162	13.956	162	13.96	72	0.2	0.077	–0.388	2	4	111	311	3
41.15094074	29.05385945	152	2.192	146	2.175	235	0.3	–0.224	0.74	0	–13	175	56	4
41.15050148	29.0625829	142	15.32	144	15.312	53	0.5	0.393	–1.351	10	15	110	387	2
41.15012403	29.0613634	167	13.527	168	13.52	79	0.2	–0.195	–0.278	0	1	195	1484	1
41.15016	29.06152	169	18.036	169	18.036	259	0	–0.215	0.26	–10	–4	119	1189	1
41.15011	29.0585	186	17.236	182	17.184	272	1.3	–0.154	2.823	0	–26	119	711	2
41.15026906	29.05745309	161	19.645	160	19.639	250	0.5	–0.075	1.063	–4	–15	175	1294	1
41.15015	29.05846	175	15.063	173	15.055	262	0.5	–0.268	1.233	–5	–11	123	746	2
41.15003919	29.06218542	191	11.127	189	11.123	279	0.3	–0.508	1.14	–30	–14	86	1896	1
41.15007333	29.06011754	185	8.458	183	8.452	272	0.3	–0.16	0.784	–10	–8	50	2136	1
41.15044	29.0576	169	15.374	166	15.363	256	0.6	–0.315	1.479	–2	–14	123	768	2
41.15026	29.05534	148	13.976	146	13.966	236	0.5	0.443	0.583	11	–1	98	494	3

No. observations	DRIFT Entry	SPD F Entry	SPD A Entry	RUD Entry	ROT Entry	RPM L Entry	Δ Entry	μ	μ - Entry 18.RBF 14-15-4	**μ-1 18.RBF 14-15-4**	μ-2 18.RBF 14-15-4	μ-3 18.RBF 14-15-4	μ-4 18.RBF 14-15-4
2	1,100000	-0,63500	2,89900	-15,0000	-31,0000	119,0000	518,000	2	2	0,000000	1,000000	0,000000	0,000000
4	0,600000	0,04100	1,10000	0,0000	-14,0000	175,0000	621,000	2	2	0,000000	1,000000	0,000000	0,000000
7	0,000000	-0,02700	-0,05800	1,0000	0,0000	196,0000	162,000	4	4	0,030353	0,000000	0,004995	0,964652
11	0,400000	0,67800	-1,39300	1,0000	29,0000	174,0000	1045,000	1	1	0,931135	0,000000	0,000587	0,068278
12	0,200000	-0,05800	-0,26200	0,0000	1,0000	99,0000	741,000	2	2	0,000000	0,998196	0,000000	0,001804
13	2,900000	-1,02000	6,79400	-35,0000	-69,0000	115,0000	293,000	3	3	0,009578	0,000000	0,866462	0,123960
14	2,900000	0,48700	-6,22800	5,0000	94,0000	174,0000	226,000	3	3	0,000000	0,000000	1,000000	0,000000
17	0,100000	0,15900	-0,26500	7,0000	3,0000	72,0000	1615,000	1	1	1,000000	0,000000	0,000000	0,000000
18	0,100000	-0,02900	0,23000	-1,0000	-6,0000	157,0000	1732,000	1	1	0,999996	0,000000	0,000004	0,000000

Figure 12 – Indexes of network trust level №18

One more idea to be highlighted relates to the question that analysis of the sensitivity of parameter factors has the tendency of Δ indicating as the most significant one for the classification using artificial neural networks. The further following proposed table of weights reflects the same idea (Table 5).

Considering the application of the classification results for the functioning of the DSS, it is safe to look upon

DOI 10.15588/1607-3274-2020-1-15

Table 5 – Weights of network №18

Weights ID	Connections 13.RBF 14-15-4	Weight values 18.RBF 14-15-4
1	LAT – hidden neuron 1	0,00
2	LON – hidden neuron 1	0,40
3	COG – hidden neuron 1	0,63
4	SOG – hidden neuron 1	0,36
5	HDG – hidden neuron 1	0,57
6	LOG – hidden neuron 1	0,36
7	SET – hidden neuron 1	0,94
3	DRIFT – hidden neuron 1	0,10
9	SPD F – hidden neuron 1	0,42
10	SPD A – hidden neuron 1	0,55
11	RUD – hidden neuron 1	0,47
12	ROT – скрытый нейрон 1	0,35
13	RPM L – hidden neuron 1	0,00
14	Δ – hidden neuron	1,00

the performing the maneuver μ_1 as being the most effective strategy to be chosen. To prove the foregoing information an erroneous judgment of the skipper during the experimental survey is turning out to be a suitable and a reliable one.

This available data indicates an implicit predisposition to a distortion of the perception of navigational danger. This conclusion can be acquired being completely confirmed by elevators cards and 3D graphs of the confidence level in the Figure 13 a, b.

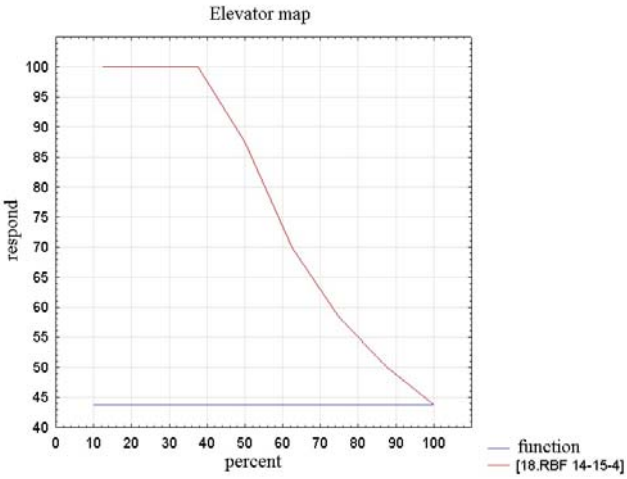
As it clearly seen, the worst result is noticed to be μ_4.

This issue being coherent and comprehensible enough can easily be clarified due to the fact that in most cases this strategy might not always be accurate and, as a result, might not be recommended for use in regular situations. To be exactly, turning with anchor drop is named to be the most effective one in emergency cases that threaten to collide with another vessel or land aground. To testify this the idea is illustrated by the proposed graphs (Fig. 14). Strategies μ_2 and μ_3 do not possess competitive nature with the given μ_1 being located in an intermediate position between μ_1 and μ_4.

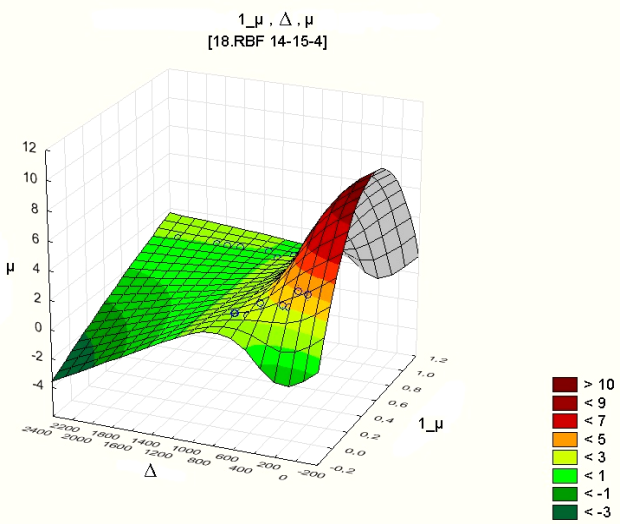
Obviously, with an aim to fulfill the DSS forecast, a broader skipper model must meet the needs to be revealed paying close attention to the question of having this issue classified by a sufficient number of locations with a target to predict its perception of behavior anywhere in the world.

It is certain enough to say that a large and long-term analysis by both the DSS and the staff of the Kheson State Maritime Academy is highly requested to be fulfilled.

Besides, these minor approaches might be dealt with and settled up apparently as they are highly likely to prevent the negative consequences of distortion in the perception of navigational danger by cadets during internships and directly during performing navigation tasks and controlling the vessel process.



a



b

Figure 13 – Confidence level chart for μ_1

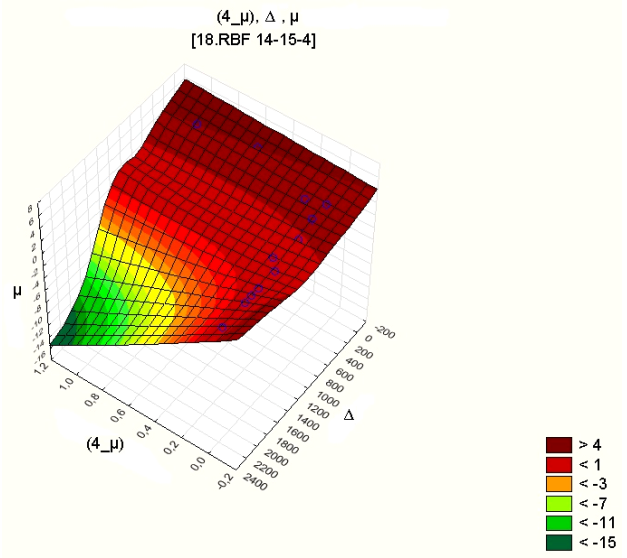


Figure 14 – Confidence level chart for μ_4

DOI 10.15588/1607-3274-2020-1-15

5 RESULTS

It is certain to underline that the approaches to identifying navigational hazards are based mostly on probabilistic models and a formal description in the framework of set theory was critically examined and analyzed. The proposed above strategy is named to be justified making it possible to bring out and set interrelation between the concepts of situation perception, location features and probable risks.

In addition there is a need to add that the applied criteria for constructing an incomplete model of perception by a operator of dangers is mostly limited by inaccuracies and subjective experimental relationships regarding the local metric of the transition of the vessel in the nearest interaction radius.

Furthermore, a DSS operation scheme is proposed to be undertaken. This item involves variety of factors to be observed that prone to influence on operator decision making, as well as a formal algorithm for test identification of operator preferences when choosing and setting a resulting strategy based on Pareto efficiency theory upper bounds for optimal alternatives.

Therefore, an experiment is announced to have been conducted on the basis of an automated analysis relating to the point of the strategies choice being accomplished by the operator when performing maneuvers based on data from logfiles for 2.5 years period. The processed data by means of artificial neural networks happened to illustrate nonrandom dependencies that form class-forming sets of parameters and to indicate in an acceptable way an erroneous perception of the situation in difficult conditions.

Basing on the following, eventually, having been received results of the experimental data processing, the proposed models and algorithms welcomed the introduction of identification process of the stable models of operator behavior in critical situations. The outcome was managed to be obtained practically through experience. Consequently, this practice provided assistance to us in operation of constructing a DSS with the ability to deliver prediction and to prevent negative consequences in maritime transport as well as to build an individual operator model in the form of a preference map.

6 DISCUSSION

The developed models and algorithms succeed in empowering to exclude the negative manifestations of stable preferences of the operator when choosing a maneuvering strategy in tight spaces.

To conclude, it could be argued that due to the results of modeling and analysis of the experimental data of the Navi Trainer 5000 navigation simulator the proposed formal approaches and software lead the way to high level determining of possible errors in perceiving dangers from the side of the operator and, as a result, to prevention of catastrophes when performing complex maneuvers.

CONCLUSIONS

It can be summarized that models and algorithms for identifying distortions in the perception of dangers by the operator during the execution of complex maneuvers are reported to be defined.

The results of the experiment using NTPRO 5000, as well as the developed software, enable us to reveal the subjective errors in making managerial decisions made by the operator. Identification is grounded on the ratio of adaptive test results based on Pareto-optimal alternatives for preference and, as for classification, it is grounded on artificial neural networks. It thus became beneficial to create the environment likely to have the criteria defined for the formation of an information model of the operator in the conditions of partial watch team and the realization of complex maneuvers.

So, the further improvement which can be strongly supported lies in the area of the development of DSS [34] software on the ground of the proposed models and algorithms for detecting deviations in the perception of navigational hazards. Basing on these investigations the level of safety of watch keeping in the condition of its incompleteness can be significantly increased.

Special emphasis must be added to the point of introducing an algorithm, which is considered to be a scientific novelty. It is the effective and reasonable way, from one point of view, to provide possibility to identify stable subjective connections that affect the adequate perception of the navigation situation and, from another one, to forecast the performance of similar maneuvers based on classification analysis.

The practical relevance of the research, in fact, is reported to depict that the results of the experiment, as well as the developed software, are drafted in such a way as to make it possible to identify preferences in the choice of strategies and their modifications as well as to separate the uncertain actions of the operator from the actions verified many times. In view of this fact, random positive maneuvering results from the side of the operator are found out to be advantageous enough being characterized by a spontaneous choice of strategy and, by the way, they do not guarantee success under similar conditions in the future. These spoken above developments are broadly looked at as being worthwhile for reinforcing the skills of cadets of naval academies during training simulations on navigation simulators.

One more issue to be paid attention to is that further proposed way of prospects is certain to be preceding to the development of methods for eliminating erroneous decision-making principles when choosing maneuvering strategies managing a vessel.

ACKNOWLEDGEMENTS

The work is carried out within the framework of “Development of software for important the quality of functioning systems for dynamic position sea vessels” (state registration number 0119U100948), of navigation and ECDIS departments of Kherson State Maritime Academy Navigation Faculty (scientific adviser: Ph.D. Associate Professor, Deputy Rector for scientific and pedagogical work, Kherson State Maritime Academy, Ukraine, Ben A.P.).

DOI 10.15588/1607-3274-2020-1-15

REFERENCES

1. Puisa R., Lin L., Bolbot V. et al. Unravelling causal factors of maritime incidents and accidents, Safety Science, 2018, Vol. 110(A), pp. 124–141. DOI: 10.1016/j.ssci.2018.08.001.
2. Tran N. K., Haasis H. D. A research on operational patterns in container liner shipping, *Transport*, 2018, Vol. 33, Issue 3, pp. 619–632. DOI: 10.3846/transport.2018.1571.
3. Topolšek D., Dragan D. Relationships between the motorcyclists' behavioural perception and their actual behavior, *Transport,* 2018, Vol. 33, Issue 1, pp. 151–164. DOI: 10.3846/16484142.2016.1141371.
4. Nosov P. S., Palamarchuk I. V., Safonov M. S. et al. Modeling the manifestation of the human factor of the maritime crew, *Science and transport progress*, 2018, Vol. 5, Issue 77, pp. 82–92. DOI:10.15802/stp2018/ 147937.
5. Pytev Yu. M. Reliability of interpretation of an experiment based on an approximate model, *Math modeling*, 1989, Vol. 1, Issue 2, pp. 49–64.
6. Stepantsov M. E. Mathematical model for the directed motion of a people group, *Math modeling*, 2004, Vol. 16, Issue 3, pp. 43–49.
7. [Nosov P. S., Ben A. P., Matejchuk V. N. et al. Identification of "Human error" negative manifestation in maritime transport, *Radio Electronics, Computer Science, Control,* 2018, Vol. 4, Issue 47, pp. 204–213. DOI: 10.15588/1607-3274-2018-4-20.
8. Bole A., Wall A., Norris A. Navigation Techniques Using Radar and ARPA, *Radar and ARPA Manual: Third Edition.* Butterworth-Heinemann, 2014, pp. 371–405. DOI: 10.1016/B978-0-08-097752-2.00008-8.
9. Zinchenko S. M., Nosov P. S., Mateichuk V. M. et al. Use of navigation simulator for development and testing ship control systems, *The international scientific and practical conference dedicated to the memory of professors Fomin Y. Y. and Semenov V. S.* Odessa. Stambul, 24–28 April 2019, proceedings. ONMU, 2019, pp. 350–355.
10. Shiqi F., Jinfen Z., Eduardo B. D. et al. Effects of seafarers emotion on human performance using bridge simulation, *Ocean Engineering*, 2018, Vol. 170, pp. 111–119. DOI: 10.1016/j.oceaneng.2018. 10.021.
11. Popovych I. S., Blynova O. Ye. The structure, variables and interdependence of the factors of mental states of expectations in students' academic and professional activities, *The New Educational Review,* 2019, Vol. 55, Issue 1, pp. 293–306. DOI: 10.15804/ tner.2019.55.1.24.
12. COLREGS – International Regulations for Preventing Collisions at Sea [Electronic resource]. Access mode: http://www.jag.navy.mil/distrib/instructions/COLREG-1972.pdf
13. Xi Y., Yang Z. , Fang Q. et al. A new hybrid approach to human error probability quantification-applications in maritime operations, *Ocean Engineering*, 2017, Vol. 138, pp. 45–54. DOI: 10.1016/j.oceaneng.2017. 04.018.
14. Akyuz E. Quantitative human error assessment during abandon ship procedures in maritime transportation, *Ocean Engineering*, 2016, Vol. 120, pp. 21–29. DOI:10.1016/j.oceaneng.2016.05. 017.
15. Park Y. A., Yip T. L., Park H. G. An Analysis of Pilotage Marine Accidents in Korea, *The Asian Journal of Shipping and Logistics*, 2019, Vol. 35, Issue 1, pp. 49–54. DOI: 10.1016/j.ajsl.2019.03.007.
16. Nosov P., Ben A., Safonova A. et al. Formal going approaches to determination periods of intuitional behavior of navigator during supernumerary situations, *Radio Electronics, Computer Science, Control*, 2019, Vol. 2, Issue 49, pp. 140–150. DOI: 10.15588/1607-3274-2019-2-15.
17. Pauer G., Sipos T., Török Á. Statistical analysis of the effects of disruptive factors of driving in simulated environment, *Transport,* 2019, Vol. 34, Issue 1, P. 1–8. DOI: 10.3846/transport. 2019.6724.
18. Szlapczynski R., Krata P. Determining and visualizing safe motion parameters of a ship navigating in severe weather conditions, *Ocean Engineering*, 2018, Vol. 158, pp. 263–274. DOI: 10.1016/ j.oceaneng. 2018.03.092.
19. Rolf J., Asbjorn B., Aalberg L. Maritime navigation accidents and risk indicators: An exploratory statistical analysis using AIS data and accident reports, *Reliability Engineering & System Safety*, 2018, Vol. 176, pp. 174–186. DOI: 10.1016/j.ress.2018. 03.033.
20. Guidance notes on safety culture and leading indicators of safety, *American Bureau of Shipping*, 2012. Houston, Vol. 74.
21. Berg H. P. Human Factors and Safety Culture in Maritime Safety, *The International Journal on Marine Navigation and Safety of Sea Transportation*, 2013, Vol. 7, Issue 3, pp. 343–352. DOI: 10.12716/1001.07.03.04.
22. Ventikos N., Papanikolaou A., Louzis K. et al. Statistical analysis and critical review of navigational accidents in adverse weather conditions, *Ocean Engineering*, 2018, Vol. 163, pp. 502–517. DOI: 10.1016/j.oceaneng. 2018.06.001.
23. [Paulauskas V., Paulauskas D., Plačienė B. et al. Ship mooring to jetties under the crosscurrent, *Transport*, 2018, Vol. 33, Issue 2, P. 454–460, DOI: 10.3846/ 16484142.2017.1354069.
24. Jech T. Set theory. Berlin, Springer, 1997, 753 p. DOI: 10.1007/3-540-44761-X.
25. Roger B. Myerson. Game Theory: Analysis of Conflict. Harvard, 1991, 600 p.
26. Bain L. Introduction to Probability and Mathematical Statistics / L. J. Bain, M. Engelhardt. – Belmont : Duxbury Press, 1992. – 648 p.
27. Popovych I. S. Social expectations – a basic component of the system of adjusting of social conduct of a person, *Australian Journal of Scientific Research,* 2014, Vol. 2, Issue 6, pp. 393–398.
28. Dinh G. H. The combination of analytical and statistical method to define polygonal ship domain and reflect human experiences in estimating dangerous area, *International Journal of e-Navigation and Maritime Economy*, 2016, Vol. 4, pp. 97–108, DOI: 10.1016/j.enavi. 2016.06.009.
29. Olijnik A. O., Skrupskij S. Yu., Subbotin S. O. et al. Planuvannya resursiv paralelnoyi obchislyuvalnoyi sistemi pri sintezi nejronechitkih modelej dlya obrobki velikih danih, *Radio Electronics, Computer Science, Control,* 2016, Vol. 4. pp. 61–69. DOI 10.15588/ 1607-3274-2016-4-8.
30. Heiko R. The complexity of Nash equilibria, local optima, and Pareto-optimal solutions: thesis Doktors der Naturwissenschaften genehmigte Dissertation. Rheinisch-Westf˙alischen, Erlangung, 2008, 171 p.
31. Prokopchuk Y. A. Sketch of the formal theory of creativity / Dnepr, PSACEA Press, 2017, 452 p.
32. Panin V. V., Doronin V. V., Spiyan O. M. Construction of a neural network expert system for navigation data processing in terms of river e-navigation, *Radio Electronics, Computer Science, Control*, 2019, Vol. 1, pp. 203–217. DOI 10.15588/1607-3274-2019-1-19.
33. De Luca M. A comparison between prediction power of artificial neural networks and multivariate analysis in road

DOI 10.15588/1607-3274-2020-1-15

safety management, *Transport,* 2017, Vol. 32, Issue 4, pp. 379–385, DOI: 10.3846/ 16484142.2014.995702.
34. Firsov S. N., Pishchukhina O. A. Intelligent support of multilevel functional stability of control and navigation systems *Radio Electronics, Computer Science, Control,* 2018, Vol. 2. pp. 177–183. DOI 10.15588/1607-3274-2018-2-20.

Received 11.09.2019.
Accepted 30.12.2019.

УДК 004.942: 316.454.54

СИСТЕМА ДІАГНОСТИКИ СПРИЙНЯТТЯ НАВІГАЦІЙНОЇ НЕБЕЗПЕКИ ПІД ЧАС ВИКОНАННЯ СКЛАДНИХ МАНЕВРІВ

Носов П. С. – канд. техн. наук, доцент кафедри судноводіння і електронних навігаційних систем, Херсонська державна морська академія, Україна.

Зінченко С. М. – канд. техн. наук, старший викладач кафедри управління судном, завідувач лабораторією електронних симуляторів, Херсонська державна морська академія, Україна.

Попович І. С. – д-р пед. наук, доцент кафедри загальної та соціальної психології, Херсонський державний університет, Україна.

Бень А. П. – канд. техн. наук, доцент, заступник ректора з науково-педагогічної роботи, Херсонська державна морська академія, Україна.

Нагрибельний Я. А. – канд. істор. наук, декан факультету судноводіння, Херсонська державна морська академія, Україна.

Матейчук В. М. – аспірант, завідувач лабораторією електронних симуляторів, Херсонська державна морська академія, Україна.

АНОТАЦІЯ

Актуальність. У статті розглядається задача автоматизованого аналізу прийняття рішень оператором в ергатичних системах критичних інфраструктур на прикладі управління морським транспортом в складних навігаційних умовах. Основним критерієм адекватного сприйняття вхідної інформації оператором є прогнозування поведінкових стратегій прийняття рішень в умовах дискретного часу. Однак, складність моделювання дій оператора полягає у нелінійному формуванні рішень в умовах позаштатних ситуацій і відхилень від Кодексів і Правил.

Мета. Метою дослідження є розробка математичного забезпечення модуля системи підтримки прийняття рішень (СППР) для ідентифікації класоутворюючих множин атомарних елементів що визначають факт спотворення сприйняття інформації про навігаційні ризики шляхом формального аналізу і прогнозу моделей поведінки оператора при управлінні судном.

Метод. З метою автоматизації аналізу сприйняття небезпеки оператором, була побудована математична модель, яка ідентифікує факт спотворення сприйняття у вигляді простору ознак метаданих, що одержуються за допомогою обробки інформації навігаційних приладів. Фактором дезорієнтації оператора також може служити несення вахти на зміщеному містку, що істотно впливає на аналіз інформації для адекватного прийняття рішень. У зв'язку з порушенням синхронізації навігаційних приладів, таких як: РЛС, ARPA, AIC, ECDIS, особливо у випадках виходу з режиму автоматичного управління, виникає небезпечний прецедент що полягає у неготовності оператора сприйняти складність ситуації, внаслідок чого проведено формальний аналіз на предмет підвищення ризиків під час переходу у вказаних умовах. Також побудована імовірнісна модель сприйняття ситуації в умовах картежа похибок. Моделювання показало, що без чіткої ідентифікації факторів, що впливають на спотворення сприйняття оператора, неможливо визначити ступінь критичності навігаційної ситуації, тому узагальнених статистичних даних недостатньо і для результативної роботи СППР, тобто необхідна індивідуальна інформаційна модель кожного оператора. З метою аналізу сприйняття інформації оператором був розроблений тест, що визначає переваги при виборі стратегії керуючих впливів у вигляді виконання маневру при складних навігаційних умовах. Результати тестування, а також дані по проходженню локації дозволили виконати класифікаційний аналіз по 15 параметрам за допомогою штучних нейронних мереж і визначити межі відхилень у сприйнятті навігаційної небезпеки. У ядро СППР внесений ряд правил і алгоритмів, які включають: поле взаємодії, засоби синхронізації РЛС і НІС; фактична навігаційна небезпека в даній картографічній області; траєкторії руху суден і, як результат, моделювання імовірного відхилення у сприйнятті оператора.

Результати. З метою підтвердження результативності розробленої СППР і запропонованих формально-аналітичних підходів був проведений експеримент із застосуванням навігаційного тренажера Navi Trainer 5000 (NTPRO 5000). Метадані експерименту за 2,5 року роботи навігаційних тренажерів і програмних засобів СППР дозволили ідентифікувати ймовірність відхилення в інформаційному сприйнятті небезпек і експортувати прогнозовані дані в нові для оператора локації і картографічні райони. Проведений експеримент підтвердив гіпотезу дослідження і показав доцільність трансформаційних змін даної СППР для виконання прогнозів можливих ризиків при управлінні судном шляхом аналізу інформаційної моделі оператора.

Висновки. Представлені в дослідженні інформаційні та формально-аналітичні підходи, а також розроблені програмні засоби СППР дозволили виконати класифікацію стратегій прийняття рішень оператором при управлінні судном і спрогнозувати ймовірність катастрофічних наслідків. Проведені експерименти підтвердили доцільність запропонованих моделей і методів.

КЛЮЧОВІ СЛОВА: системи підтримки прийняття рішень; інформаційна модель оператора; комп'ютерні навігаційні симулятори; ймовірність ризиків; інформаційний аналіз людського фактора, автоматизовані системи управління, автоматичні системи управління.

DOI 10.15588/1607-3274-2020-1-15

УДК 004.942:316.454.54

СИСТЕМА ДИАГНОСТИКИ ВОСПРИЯТИЯ НАВИГАЦИОННОЙ ОПАСНОСТИ ПРИ ВЫПОЛНЕНИИ СЛОЖНЫХ МАНЕВРОВ

Носов П. С. – канд. техн. наук, доцент кафедры судовождения и електронных навигационных систем, Херсонская государственная морская академия, Украина.

Зинченко С. Н. – канд. техн. наук, старший преподаватель кафедры управления судном, заведующий лабораторией электронных симуляторов, Херсонская государственная морская академия, Украина.

Попович И. С. – д-р пед. наук, доцент кафедры общей и социальной психологии, Херсонский государственный университет, Украина.

Бень А. П. – канд. техн. наук, доцент, заместитель ректора по научно-педагогической работе, Херсонская государственная морская академия, Украина.

Нагрибельный Я. А. – канд. истор. наук, декан факультета судовождения, Херсонская государственная морская академия, Украина.

Матейчук В. Н. – аспирант, заведующий лабораторией электронных симуляторов, Херсонская государственная морская академия, Украина.

АННОТАЦИЯ

Актуальность. В статье рассматривается задача автоматизированного анализа принятия решений оператором в эргатических системах критических инфраструктур на примере управления морским транспортом в сложных навигационных условиях. Основным критерием адекватного восприятия входной информации оператором является прогнозирование поведенческих стратегий принятия решений в условиях дискретного времени. Однако, сложность моделирования действий оператора состоит в нелинейном формировании решений в условиях внештатных ситуаций и отклонений от Кодексов и Правил.

Цель. Целью исследования является разработка математического обеспечения модуля системы поддержки принятия решений (СППР) для идентификации классо-образующего множества атомарных элементов, определяющих факт искажения восприятия информации о навигационных рисках путем формального анализа и прогноза моделей поведения оператора при управлении судном.

Метод. С целью автоматизации анализа восприятия опасности оператором, была построена математическая модель, которая идентифицирует факт искажения восприятия в виде признакового пространства метаданных, получаемых посредством преобразования информации навигационных приборов. Фактором дезориентации оператора также может служить несение вахты на смещенном мостике, что существенно влияет на анализ информации для адекватного принятия решений. В связи с нарушением синхронизации навигационных приборов, таких как: РЛС, ARPA, AIC, ECDIS, особенно в случаях выхода из режима автоматического управления, создается опасный прецедент неготовности оператора воспринять сложность ситуации вследствие чего проведен формальный анализ на предмет повышения рисков во время перехода в этих условиях. Также построена вероятностная модель восприятия ситуации в условиях картежа погрешностей. Моделирование показало, что без четкой идентификации факторов, влияющих на искажение восприятия оператора, невозможно определить степень критичности навигационной ситуации, поэтому обобщенных статистических данных недостаточно и для результативной работы СППР необходима индивидуальная информационная модель каждого оператора. С целью анализа восприятия информации оператором был разработан тест, определяющий предпочтения при выборе стратегии управляющих воздействий в виде выполнения маневра при сложных навигационных условиях. Результаты тестирования, а также данные по прохождению локаций позволили выполнить классификационный анализ по 15 параметрам с помощью искусственных нейронных сетей и определить границы отклонений в восприятии навигационной опасности. В ядро СППР внесен ряд правил и алгоритмов, включающие: поле взаимодействия, средства синхронизации РЛС и НИС; фактическая навигационная опасность в данной картографической области; траектории движения судов и, как результат, моделирования вероятного отклонения в информационном восприятии оператора.

Результаты. С целью подтверждения результативности разработанной СППР и предложенных формально-аналитических подходов был проведен эксперимент с применением навигационного тренажера Navi Trainer 5000 (NTPRO 5000). Метаданные эксперимента за 2,5 года работы навигационных тренажеров и программные средства СППР позволили идентифицировать вероятность отклонения в информационном восприятии опасностей и экспортировать прогнозированные данные в новые для оператора локации и картографические районы. Проведенный эксперимент подтвердил гипотезу исследования и показал целесообразность применения данной СППР для выполнения прогнозов возможных рисков при управлении судном путем анализа информационной модели оператора.

Выводы. Представленные в исследовании информационные и формально-аналитические подходы, а также разработанные программные средства СППР позволили выполнить классификацию стратегий принятия решений оператором при управлении судном и спрогнозировать вероятности катастрофических последствий. Проведенные эксперименты подтвердили целесообразность предложенных моделей и методов.

КЛЮЧЕВЫЕ СЛОВА: системы поддержки принятия решений; информационная модель оператора; компьютерные навигационные симуляторы; вероятность рисков; информационный анализ человеческого фактора, автоматизированные системы управления, автоматические системы управления.

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Unravelling causal factors of maritime incidents and accidents / [R. Puisa, L. Lin, V. Bolbot et al.] // Safety Science. – 2018. – Vol. 110(A). – P. 124–141. DOI: 10.1016/j.ssci. 2018.08.001.
2. Tran N. K. A research on operational patterns in container liner shipping / N. K. Tran, H. D. Haasis // Transport. – 2018. – Vol. 33, Issue 3. – P. 619–632. DOI: 10.3846/transport. 2018.1571.
3. Topolšek, D. Relationships between the motorcyclists' behavioural perception and their actual behavior / D. Topolšek,

DOI 10.15588/1607-3274-2020-1-15

D. Dragan // Transport. – 2018. – Vol. 33, Issue 1. – P. 151–164. DOI: 10.3846/16484142.2016.1141371.
4. Modeling the manifestation of the human factor of the maritime crew / [P. S. Nosov, I. V. Palamarchuk, M. S. Safonov et al.] // Science and transport progress. – 2018. – Vol. 5, Issue 77. – P. 82–92. DOI:10.15802/stp2018/ 147937.
5. Пытьев Ю. М. Надежность интерпретации эксперимента, основанной на приближенной модели / Ю. М. Пытьев // Математическое моделирование. – 1989. – Том 1, Номер 2. – С. 49–64.
6. Степанцов М. Е. Математическая модель направленного движения группы людей / М. Е. Степанцов // Математическое моделирование. – 2004. – Том 16, Номер 3. – С. 43–49.
7. Identification of "Human error" negative manifestation in maritime transport / [P. S. Nosov, A. P. Ben, V. N. Matejchuk et al.] // Radio Electronics, Computer Science, Control. – 2018. – Vol. 4, Issue 47. – P. 204–213. DOI: 10.15588/1607-3274-2018-4-20.
8. Bole A. Navigation Techniques Using Radar and ARPA / A. Bole // Radar and ARPA Manual: Third Edition / A. Bole, A. Wall, A. Norris. – Butterworth-Heinemann, 2014. – P. 371–405. DOI: 10.1016/B978-0-08-097752-2.00008-8.
9. Use of navigation simulator for development and testing ship control systems / [S. M. Zinchenko, P. S. Nosov, V. M. Mateichuk at al.] // The international scientific and practical conference dedicated to the memory of professors Fomin Y. Y. and Semenov V. S., Odessa – Stambul, 24–28 April 2019, proceedings. – ONMU, 2019. – P. 350–355.
10. Effects of seafarers emotion on human performance using bridge simulation / [F. Shiqi, Z. Jinfen, B. D. Eduardo at al.] // Ocean Engineering. – 2018. – Vol. 170. – P. 111–119. DOI: 10.1016/j.oceaneng.2018. 10.021.
11. Popovych I. S. The structure, variables and interdependence of the factors of mental states of expectations in students' academic and professional activities / I. S. Popovych, O. Ye. Blynova // The New Educational Review. – 2019. – Vol. 55, Issue 1. – P. 293–306. DOI:10.15804/ tner.2019.55.1.24.
12. COLREGS – International Regulations for Preventing Collisions at Sea [Electronic resource]. – Access mode: http://www.jag.navy.mil/distrib/instructions/COLREG-1972.pdf
13. A new hybrid approach to human error probability quantification–applications in maritime operations / [Y. Xi, Z. Yang, Q. Fang et al.] // Ocean Engineering. – 2017. – Vol. 138. – P. 45–54. DOI: 10.1016/j.oceaneng.2017. 04.018.
14. Akyuz E. Quantitative human error assessment during abandon ship procedures in maritime transportation / E. Akyuz // Ocean Engineering. – 2016. – Vol. 120. – P. 21–29. DOI:10.1016/j.oceaneng.2016.05. 017.
15. Park Y. A. An Analysis of Pilotage Marine Accidents in Korea / Y. A. Park, T. L. Yip, H. G. Park // The Asian Journal of Shipping and Logistics. – 2019. – Vol. 35, Issue 1. – P. 49–54. DOI: 10.1016/j.ajsl.2019.03.007.
16. Formal going approaches to determination periods of intuitional behavior of navigator during supernumerary situations / [P. Nosov, A. Ben, A. Safonova at al.] // Radio Electronics, Computer Science, Control. – 2019. – Vol. 2, Issue 49. – P. 140–150. DOI: 10.15588/1607-3274-2019-2-15.
17. Pauer G. Statistical analysis of the effects of disruptive factors of driving in simulated environment / G. Pauer, T. Sipos, Á. Török // Transport. – 2019. – Vol. 34, Issue 1. – P. 1–8. DOI: 10.3846/transport. 2019.6724.
18. Szlapczynski R. Determining and visualizing safe motion parameters of a ship navigating in severe weather conditions / R. Szlapczynski, P. Krata // Ocean Engineering. – 2018. – Vol. 158. – P. 263–274. DOI: 10.1016/ j.oceaneng. 2018.03.092.
19. Rolf J. Maritime navigation accidents and risk indicators: An exploratory statistical analysis using AIS data and accident reports / J. Rolf, B. Asbjorn, L. Aalberg // Reliability Engineering & System Safety. – 2018. – Vol. 176. – P. 174–186. DOI: 10.1016/j.ress.2018. 03.033.
20. Guidance notes on safety culture and leading indicators of safety / American Bureau of Shipping. – 2012. – Houston. – Vol. 74.
21. Berg H. P. Human Factors and Safety Culture in Maritime Safety / H. P. Berg // The International Journal on Marine Navigation and Safety of Sea Transportation. – 2013. – Vol. 7, Issue 3. – P. 343–352. DOI: 10.12716/1001.07.03.04.
22. Statistical analysis and critical review of navigational accidents in adverse weather conditions / [N. Ventikos, A. Papanikolaou, K. Louzis at al.] // Ocean Engineering. – 2018. – Vol. 163. – P. 502–517. DOI: 10.1016/j.oceaneng. 2018.06.001.
23. 23. Ship mooring to jetties under the crosscurrent / [V. Paulauskas, D. Paulauskas, B. Plačienė at al.] // Transport. – 2018. Vol. 33, Issue 2. – P. 454–460, DOI: 10.3846/16484142. 2017.1354069.
24. Jech T. Set theory / T. Jech. – Berlin : Springer, 1997. – 753 p. DOI: 10.1007/3-540-44761-X.
25. Roger B. Game Theory: Analysis of Conflict / B. Roger, Myerson. – Harvard, 1991. – 600 p.
26. Bain L. Introduction to Probability and Mathematical Statistics / L. J. Bain, M. Engelhardt. – Belmont: Duxbury Press, 1992. – 648 p.
27. Popovych I. S. Social expectations – a basic component of the system of adjusting of social conduct of a person / I. S. Popovych // Australian Journal of Scientific Research. – 2014. Vol. 2, Issue 6. – P. 393–398.
28. Dinh G. H. The combination of analytical and statistical method to define polygonal ship domain and reflect human experiences in estimating dangerous area / G. H. Dinh, N. K. Im // International Journal of e-Navigation and Maritime Economy. – 2016. – Vol. 4. – P. 97–108. DOI: 10.1016/j.enavi. 2016.06.009.
29. Планування ресурсів паралельної обчислювальної системи при синтезі нейро-нечітких моделей для обробки великих даних / [А. О. Олійник, С. Ю. Скрупський, С. О. Субботін та ін.] // Radio Electronics, Computer Science, Control. – 2016. – Том. 4. – С. 61–69. DOI 10.15588/ 1607-3274-2016-4-8
30. Heiko R. The complexity of Nash equilibria, local optima, and Pareto-optimal solutions: thesis Doktors der Naturwissenschaften genehmigte Dissertation / R¨oglin Heiko. – Rheinisch – Westf¨alischen: Erlangung, 2008. – 171 p.
31. Прокопчук Ю. А. Набросок формальной теории творчества / Ю. А. Прокопчук.– Днепр : ГВУЗ «ПГАСА», 2017. – 452 с.
32. Панін В. В. Побудова нейромережевої експертної системи обробки навігаційних даних в умовах річкової е-навігації / В. В. Панін, В. В. Доронін, О. М. Спіян // Радіоелектроніка, інформатика, управління. – 2019. – Том. 1. – С. 203–217. DOI 10.15588/1607-3274-2019-1-19.
33. De Luca M. A comparison between prediction power of artificial neural networks and multivariate analysis in road safety management / M. De Luca // Transport. – 2017. –Vol. 32, Issue 4. – P. 379–385, DOI: 10.3846/ 16484142.2014.995702.
34. Firsov S. N. Intelligent support of multilevel functional stability of control and navigation systems / S. N. Firsov, O. A. Pishchukhina // Радіоелектроніка, інформатика, управління. – 2018. – Vol. 2. – P. 177–183. DOI 10.15588/1607-3274-2018-2-20.

DOI 10.15588/1607-3274-2020-1-15