ITHEA

International Journal

INFORMATION MODELS & ANALYSES

2013 Volume 2 Number 2

International Journal
INFORMATION MODELS & ANALYSES
Volume 2 / 2013, Number 2

IJ IMA is official publisher of the scientific papers of the members of the ITHEA® International Scientific Society

IJ IMA rules for preparing the manuscripts are compulsory.
The **rules for the papers** for ITHEA International Journals are given on www.ithea.org .
The camera-ready copy of the paper should be received by ITHEA® Submission system http://ij.ithea.org .
Responsibility for papers published in IJ IMA belongs to authors.

General Sponsor of IJ IMA is the **Consortium FOI Bulgaria** (www.foibg.com).

International Journal "INFORMATION MODELS AND ANALYSES" Vol.2, Number 2, 2013

Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with
Institute of Mathematics and Informatics, BAS, Bulgaria,
V.M.Glushkov Institute of Cybernetics of NAS, Ukraine,
Universidad Politechnika de Madrid, Spain,
Hasselt University, Belgium
Institute of Informatics Problems of the RAS, Russia,
St. Petersburg Institute of Informatics, RAS, Russia
Institute for Informatics and Automation Problems, NAS of the Republic of Armenia,
and Federation of the Scientific - Engineering Unions /FNTS/ (Bulgaria).

Publisher: **ITHEA®**
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: info@foibg.com
Technical editor: **Ina Markova**

Printed in Bulgaria

ISSN 1314-6416 (printed) **ISSN 1314-6424 (CD)** **ISSN 1314-6432 (Online)**

A COMPARISON OF SOME APPROACHES TO THE RECOGNITION PROBLEMS IN CASE OF TWO CLASSES

Yurii I. Zhuravlev, Yuryi Laptin, Alexander Vinogradov, Aleksey Likhovid

Abstract: *We consider an improved model of the empirical risk minimization problem and its continuous relaxation. The continuous relaxation of the formulated problem is compared with the mathematical model used in the support vectors method. The results of numerical experiments comparing different models for problems with linearly inseparable sets are presented.*

Keywords: *cluster, decision rule, discriminant function, linear and nonlinear programming, nonsmooth optimization*

ACM Classification Keywords: *G.1.6 Optimization - Gradient methods, I.5 Pattern Recognition; I.5.2 Design Methodology - Classifier design and evaluation*

Introduction

Mathematical models of problems of constructing linear and non-linear classifiers and methods of constructing, based on these models, have been considered in many papers (see, e.g. [1-3]). In the present time the method of support vectors machine (SVM) is the most widely used.

For such problems it is convenient to represent mathematical models in the form of convex optimization problems. In [7] the technique using effective methods of non-smooth optimization for solving these problems was considered. The results of computational experiments were given for special large-scale test problems with linearly separable sets. A comparison was carried out with well-known program implementation LIBSVM of the method of support vector machine.

In this paper the models and approaches proposed in [4, 5] are further developed. We formulate an improved model of the empirical risk minimization problem and its continuous relaxation. The possibilities and complexity of the development of approximation algorithms to minimize the empirical risk are discussed. The continuous relaxation of the formulated problem is compared with the mathematical model used in the support vectors method. The results of numerical experiments comparing different models for problems with linearly inseparable sets are presented.

1. A brief description of problems of constructing classifiers

Let there be given a linear function $f(x,W)=\langle w,x\rangle+w_0$, where $x\in R^n$ is a vector of features, $W=(w,w_0)\in R^{n+1}$ is a vector of parameters. Function $a(x,W)$ of the following form is called *linear classifier*:

$$a(x,W)=\begin{cases}1, & \text{if } f(x,W)>0,\\ 2, & \text{if } f(x,W)\le 0.\end{cases} \qquad (2)$$

Classifier $a(x,W)$ refers each point $x\in R^n$ to one of the two classes of $\{1,\ 2\}$.

Consider a set of finite non-overlapping sets (training sample) that consists of points of R^n : $\Omega_i = \left\{ x^t : \ t \in T_i \right\}$, $i = 1,2$, $T = T_1 \bigcup T_2$.

The problem of constructing (training) classifier $a(x,W)$ is to determine the values of the parameters W based on the training sample Ω_i , $i = 1,2$.

It is said that the classifier $a(x,W)$ *correctly separates* the points of Ω_i , $i = 1,2$, if $a(x,W) = i$ for all $x \in \Omega_i$, $i = 1,2$. Classifier *gap* at a point x^t is the following value

$$g^t(W) = \begin{cases} f(x^t,W), \text{ if } t \in T_1, \\ -f(x^t,W), \text{ if } t \in T_2. \end{cases} \tag{4}$$

The value $g(W) = \min\left\{ g^t(W) : t \in T \right\}$ is called a gap of classifier $a(x,W)$ on the collection of sets Ω_i , $i = 1,2$. Classifier $a(x,W)$ correctly separates the points of the sets Ω_i , $i = 1,2$ if $g(W) > 0$.

The sets Ω_i , $i = 1,2$ are called separable in the class of linear classifiers, if there is a linear classifier, correctly separating the points of these sets.

Classifier $a(x,W)$ is invariant with respect to the multiplication function f (vector W) by a positive number, the gap $g(W)$ is linear with respect to this multiplication. The value of $g(W)$ can be used as a quality criterion for classifier $a(x,W)$ (the larger the value of $g(W)$, the more reliable the points of Ω_i , $i = 1,2$ are separeted), but we must also take into account some normalization of the vector W , which we denote $\eta(W)$ and will call the norm of the classifier $a(x,W)$.

We consider the problem of constructing an optimal classifier (determination of the values of parametres W) for the sets, Ω_i , $i = 1,2$, which are separable in the class of linear classifiers, in the following form: to find

$$g^* = \max_W \left\{ g(W) : \eta(W) \le 1, \ W \in R^{n+1} \right\} \tag{5}$$

As the norm of the vector W we use the function $\eta(W) = \sqrt{\sum_{j=1}^{n} (w_j)^2}$.

Problem (5) can be rewritten in the equivalent form

$$\eta^* = \min_V \left\{ \eta(V) : \ g(V) \ge 1, \ V \in R^L \right\} \tag{6}$$

$$\eta^* = \min_V \left\{ \eta(V) : \ g^t(V) \ge 1, t \in T, V \in R^L \right\} \tag{7}$$

This equivalence is understood in the sense that if W^* is an optimal solution of problem (5), then for optimal solutions V^* of (6) or (7) the equalities $V^* = W^* / g^*$, $\eta^* = 1 / g^*$ are satisfied [8]. Note that $g^* > 0$ for the sets which are separable in the class of linear classifiers.

2. Minimization of Empirical Risk

In the case of linearly inseparable samples the natural criterion for the choice of the classifier is the minimization of the empirical risk, i.e. number of points of training sample which the classifier separates incorrectly.

We assume that parameter $\delta > 0$ of the reliability of separating points of training sample Ω_i, $i = 1,2$ is given. The points x^t, $t \in T$ are separated by classifier $a(x,W)$ unreliably if the gap $g^t(W) < \delta$. Empirical risk with the reliability [5], defined by parameter δ, equals to the number of points of training sample, which the classifier separates incorrectly or unreliably.

The problem under consideration is to determine the minimum number of points which should be excluded from the training sample that the remaining points are separated reliably. It is natural to require that after excluding in each class at least one point is remained. This is possible if

$$\delta < \max\left\{\left\|x^\tau - x^s\right\| : \tau \in T_1,\ s \in T_2\right\} \tag{8}$$

Further we will assume that this condition is valid. It can be shown that there are sufficiently large positive numbers B_t, $t \in T$ (in [5] it was assumed that all B_t are the same) for which the empirical risk minimization problem with the reliability can be represented as the following: to find

$$Q^* = \min_{w,y}\left\{\sum_{t\in T} y_t\right\} \tag{9}$$

subject to constraints

$$g^t(W) \geq \delta - B_t \cdot y_t,\ \ t \in T \tag{10}$$

$$\langle w, w\rangle \leq 1 \tag{11}$$

$$\sum_{t\in T_i} y_t \leq |T_i| - 1,\ \ i = 1,2 \tag{12}$$

$$0 \leq y_t \leq 1,\ t \in T \tag{13}$$

$$y_t \in \{0,1\},\ t \in T \tag{14}$$

Variable y_t determines whether a point x^t is taken into account in the formulation of the problem. We say that numbers B_t, $t \in T$ satisfy the **correctness condition** if in case of $y_t = 1$ the point x^t is excluded from the training sammple, i.e. constraints (10) are satisfied for all feasible values of the other variables of the problem. Constraints (12) define the condition that at least one point from each set Ω_i should be included in the problem.

The problem (9) - (14) is NP-complete. In this regard, approximate algorithms for solving such problem must be developed for practical use. For small values of the problem dimension the existing general purpose optimization software can be used (the possibility of such approach will be considered in Section 4).

As approximate algorithms one can consider the algorithms based on the ideas of directed enumeration (sequential analysis of variants, the branch and bound methods), local search methods. Developing such algorithms it is essential to have effective procedures for calculating lower bounds for Q^* and the construction of feasible solutions of the problem (9)-(14). To implement these procedures we will use continuous relaxation of (9)-(14). It is clear that all integer formulations of the problem (9)-(14) for sufficiently large values B_t (satisfying the correctness condition) are equivalent. However, the continuous relaxation of the problem and the value of the lower bound for Q^* essentially depend on the values of B_t, since with increasing B_t the range of feasible solutions of continuous relaxation of the problem (9)-(14) is expanding. To obtain the best estimate for Q^* you must use the lowest possible values for B_t.

Let $t \in T$, $s \in T_1$, $\tau \in T_2$, $s, \tau \neq t$. Consider the problem

$$\beta_t^{s\tau} = \max\left\{\delta - g^t(W)\right\} \tag{15}$$

$$g^j(W) \geq \delta,\ j = s, \tau \tag{16}$$

$$\langle w, w \rangle \leq 1 \tag{17}$$

Denote

$$B_t^* = \max\left\{\beta_t^{s\tau} : s \in T_1, \tau \in T_2,\ s, \tau \neq t\right\},\ t \in T \tag{18}$$

Theorem 1. *Numbers* B_t, $t \in T$ *satisfy the correctness condition for problem (9) - (14) if*

$$B_t \geq B_t^*,\ t \in T \tag{19}$$

Proof. Let an index $t \in T$ be fixed. The point x^t is excluded from training sample in case of $y_t = 1$ when the constraint (10) for this index is valid for any feasible values of the remaining variables.

Denote $y = (y_\tau,\ \tau \in T)$, Y - the set of all y satisfying the constraints (12), (14), $D(y)$ - the set of all vectors W satisfying the constraints (10) and (11) for a given value of vector y. Consider the vector $y \in Y$ such that $y_t = 1$. Let

$$\beta_t(y) = \min\left\{\theta : g^t(W) \geq \delta - \theta,\ W \in D(y)\right\} = \max\left\{\delta - g^t(W) : W \in D(y)\right\}.$$

Denote $\beta_t^* = \max\left\{\beta_t(y) : y \in Y,\ y_t = 1\right\}$. It is evident that the inequality $B_t \geq \beta_t^*$ is the condition of exclusion of the point x^t from the training sample when $y_t = 1$. Let $s \in T_1$, $\tau \in T_2$, $s, \tau \neq t$. Denote $y^{s\tau} = \left(y_t, t \in T,\ y_s = 0,\ y_\tau = 0,\ y_j = 1,\ j \neq s, \tau\right)$. It is easy to see that for any $y \in Y$ such that $y_s = 0,\ y_\tau = 0$ $D(y) \subseteq D(y^{s\tau})$ is performed, i.e. $\beta_t(y) \leq \beta_t(y^{s\tau})$. Hence $\beta_t^* = \max\left\{\beta_t(y^{s\tau}) : s \in T_1, \tau \in T_2,\ s, \tau \neq t\right\}$. Taking into account that $\beta_t(y^{s\tau}) = \beta_t^{s\tau}$, i.e. $B_t^* = \beta_t^*$, we obtain the statement of the theorem.

Let $t \in T_1$. Consider in more detail the problem (15) - (17).Taking into account (4), we can rewrite this problem as

$$\beta_t^{s\tau} = -\min_{w, w_0}\left\{\left\langle w, x^t \right\rangle + w_0 - \delta\right\} \tag{20}$$

$$\left\langle w, x^s \right\rangle + w_0 \geq \delta,\ s \in T_1 \tag{21}$$

$$-\left\langle w, x^\tau \right\rangle - w_0 \geq \delta,\ \tau \in T_2 \tag{22}$$

$$\langle w, w \rangle \leq 1 \tag{23}$$

If the system of constraints (21) - (23) is inconsistent, then $\beta_t^{s\tau} = -\infty$. This occurs if $\delta > \left\| x^s - x^\tau \right\|$. By (8) there is always a pair s, τ such that $\delta \leq \left\| x^s - x^\tau \right\|$.

It is easy to see that in the optimal solution of problem (20) - (23) constraints (21), (23) must be satisfied as

equality, and constraint (22) can be either active or inactive. Consider the case when the constraint (22) is inactive in the optimal solution. Using the Lagrange multiplier rule, we obtain for optimal solutions $w = \frac{x^s - x^t}{\left\| x^s - x^t \right\|}$, $w_0 = \delta - \left\langle w, x^s \right\rangle$, $\beta_t^{s\tau} = \left\| x^s - x^t \right\|$. For the resulting vector $(w,\ w_0)$ constraint (22) should be satisfied. If this constraint is not satisfied, then the optimal solution should constructed on the fact that the constraint (22) is active. The obtained relations allow relatively easy to determine the values of B_t^*, $t \in T$.

Consider the problem (9)-(13) - the continuous relaxation of the problem of minimization of the empirical risk. Denote $d^t(W) = \max\left(0, \frac{1}{B_t}\left(\delta - g^t(W)\right)\right)$ and fix some values of the variables W. It is easy to see that if for these values W a solution of problem (9) - (13) exists, then $y^t = d^t(W)$. Hence we obtain the problem of minimization in the variables W: to find

$$q^* = \min_W \sum_{t \in T} d^t(W) \tag{24}$$

subject to

$$\left\langle w, w \right\rangle \le 1 \tag{25}$$

$$\sum_{t \in T_i} d^t(W) \le \left| T_i \right| - 1,\ i = 1, 2 \tag{26}$$

$$d^t(W) \le 1,\ t \in T \tag{27}$$

Value q^* is a lower bound for the minimum value of the empirical risk Q^* and the vector W obtained by solving the problem (24) - (27) defines an approximate solution of the problem (9) - (14). $d^t(W)$ - convex piecewise-linear functions. To solve the problem (24) - (27) it is appropriate to use effective methods of non-smooth optimization [6].

3. Method of Support Vector

In the method of support vectors (SVM) for the case $m = 2$ the following problem is solved: to find

$$\eta^* = \min_{v, v_0} \left\{ \left\langle v, v \right\rangle + C \sum_{t \in T} \xi^t \right\} \tag{28}$$

subject to

$$\left\langle v, x^t \right\rangle + v_0 \ge 1 - \xi^t,\ t \in T_1 \tag{29}$$

$$-\left\langle v, x^t \right\rangle - v_0 \ge 1 - \xi^t,\ t \in T_2 \tag{30}$$

$$\xi^t \ge 0, t \in T \tag{31}$$

The method of support vector (SVM) is used for finding an optimal classifier for linearly separable classes, and also for the classes which are linearly inseparable.

Note that constraints (29) and (30) correspond to the constraint $g^t(V) \ge 1,\ \ t \in T$. In the case of linearly separable classes it follows from theorems of non-smooth penalties (see, for example, [6]) that for a sufficiently

large value of C the problems (7) and (28) - (31) have the same solution. In the case of linearly inseparable classes the problem (28) - (31) is interpreted as some regularization of the empirical risk minimization problem. We will show that there are certain relationships between the problem (28) - (31) and the continuous relaxation (9) - (13) of empirical risk minimization problem.

Relax the constraints (10), setting $B_t = B := \max\limits_{\tau} B_\tau^*$ and exclude the constraint (12). We obtain the following problem:

$$\overline{q}^* = \min_{w,y} \left\{ \sum_{t \in T} y_t \right\} \tag{32}$$

subject to

$$\left\langle w, x^t \right\rangle + w_0 \geq \delta - B \cdot y_t, \ t \in T_1 \tag{33}$$

$$-\left\langle w, x^t \right\rangle - w_0 \geq \delta - B \cdot y_t, \ t \in T_2 \tag{34}$$

$$\langle w, w \rangle \leq 1 \tag{35}$$

$$y_t > 0, \ t \in T \tag{36}$$

Here, the constraint (11) is replaced by the equivalent pair of constraints (33) and (34). It is clear that $\overline{q}^* \leq q^*$. Let make a change of the variables $w = \delta v, \ w_0 = \delta v_0$, $\xi^t = \dfrac{B y_t}{\delta}$, $t \in T_1 \bigcup T_2$. The problem takes the form

$$\overline{q}^* = \frac{\delta}{B} \cdot \min_{v, v_0, \xi} \left\{ \sum_{t \in T} \xi^t \right\} \tag{37}$$

subject to

$$\left\langle v, x^t \right\rangle + v_0 \geq 1 - \xi^t, \ t \in T_1 \tag{38}$$

$$-\left\langle v, x^t \right\rangle - v_0 \geq 1 - \xi^t, \ t \in T_2 \tag{39}$$

$$\langle v, v \rangle \leq 1 \Big/ \delta^2 \tag{40}$$

$$\xi_t > 0, \ t \in T \tag{41}$$

Let $\alpha \geq 0$ be a dual variable for constraint (40). Consider the Lagrangian function $L(\alpha, \xi, v) = \dfrac{\delta}{B} \sum\limits_{t \in T} \xi^t + \alpha \cdot (\langle v, v \rangle - 1 \Big/ \delta^2)$ and Lagrangian relaxations of the problem (37) - (41): to find

$$\varphi(\alpha) = \min_{v, v_0, \xi} L(\alpha, \xi, u) \tag{42}$$

subject to (38), (39), (41).

Since $\varphi(\alpha)$ is the optimal value of the Lagrangian relaxation of (37) - (41), then $\varphi(\alpha) \leq \overline{q}^*$ for any $\alpha \geq 0$ (see, e.g., [6]). Given a penalty factor C in (28) - (31), choose α from the condition $\dfrac{\delta}{\alpha B} = C$. We obtain

$$L(\alpha, \xi, u) = \alpha \left\{ \langle v, v \rangle + C \cdot \sum_{t \in T} \xi^t \right\} - \frac{\alpha}{\delta^2},$$

i.e. the problem (42), (38), (39), (41) is equivalent to (28) - (31) with accuracy to an additive constant and a fixed factor in the objective function value for the above choice of the dual variable.

Thus, the problem (28) - (31), which is solved by the method of support vectors can be obtained as a result of relaxing constraints of (24) - (27), which in turn is a continuous relaxation of the problem of minimization of the empirical risk.

4. The results of numerical experiments

The quality of solutions obtained by using the empirical risk minimization model (9) - (14), the continuous relaxation of (9) - (13) and the model SVM (28) - (31) is compared in the computational experiments. Quality criterion is the error of classification - the number of training sample points that are classified incorrectly. The well-known software package CPLEX is used to solve the generated problems. Points in the training sample for each class were generated on the basis of a uniform distribution in the unit cube. These cubes are shifted relative to each other in the first coordinate so that the distance between them is 0.1. Family of linearly inseparable sets is formed iteratively, by moving at the current iteration a single point of each class to the opposite one.

Model (9) - (14) is NP-complete, so problems of low dimension were generated for numerical experiments. Fig. 1 shows the results for the case when $|\Omega_i| = 25,\ i = 1, 2$, ($\Omega_i$ - points of the training sample for the class i) n = 5 (n - dimension of the feature space R^n). For 25 iterations all points of a class move to the other, and vice versa. On X-axis the number of moved points of a class is indicated, the vertical axis - the classification error, MER - empirical risk minimization model (9) - (14), RMER - the relaxed model of minimization of the empirical risk (9) - (13). The complexity of the exact solution of the empirical risk minimization problems (9) - (14) for the family, shown in Figure 1, reached 90 min. Solving problems of larger dimension we obtained the messages of the package CPLEX for failure of computing resources.

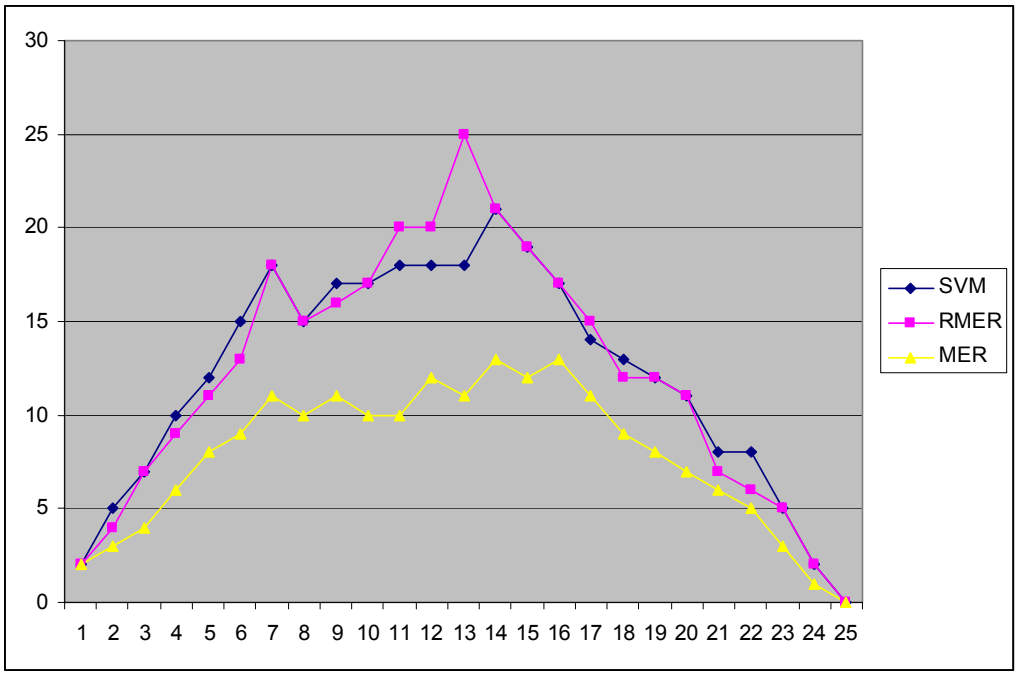


Figure 1. The dependence of the classification error on the number of displaced points n = 5, $|\Omega_i| = 25,\ i = 1, 2$.

In this regard, the comparison for the large-scale problems was realized only for the relaxed model of minimization of the empirical risk RMER and model SVM. Fig. 2 shows the results for the case $|\Omega_i| = 100,\ i = 1,2$, n = 30. It is essential to analyze the possibilities of the different models for the problems in which the value $|\Omega_i|,\ i = 1,2$ are significantly different. For this case it is necessary to estimate the value of the error of classification separately for each class. Fig. 3 shows the results for the case $|\Omega_1| = 30,\ |\Omega_2| = 200$, n = 30. The number of iterations for constructing a family of problems is 30.

Conclusion

The paper discusses various approaches to solving the problems of classification in the case of two classes. For linearly inseparable sets a mixed-integer model of the problem of minimization of the empirical risk and the continuous relaxation of the model are considered. It is shown that at weakened constraints of the proposed continuous relaxation the mathematical model used in the method of support vectors can be obtained.

The results of numerical experiments comparing approaches considered for the case of linearly inseparable sets are given. Classification error obtained by using the model of minimization of the empirical risk is much smaller than the error obtained when using continuous relaxation of this model and SVM method. This comparison was made for the problems of low dimension due to NP-completeness of the first model. Comparison of the second and the third models was also performed for large-scale problems. The resulting classification errors were about the same.

From the obtained results one can make a conclusion that it is appropriate to develop the approximate algorithms for solving the problem of minimization of the empirical risk based on the ideas of directed enumeration (sequential analysis of variants, branch-and-bound methods), local search methods, to improve the quality of generated classifiers.

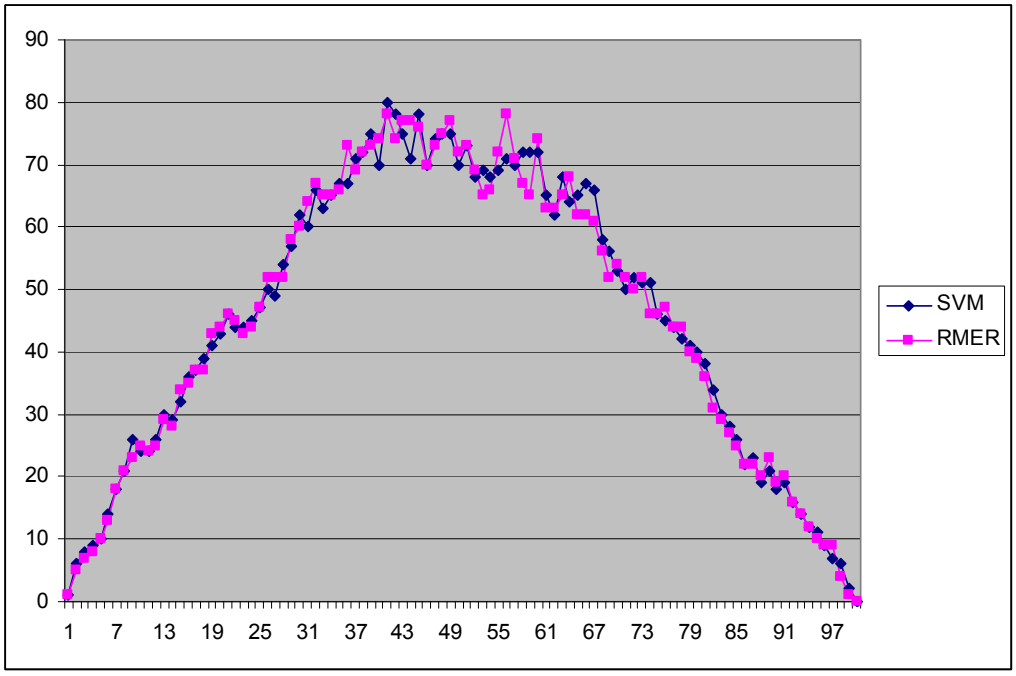


Figure 2. The dependence of the classification error on the number of displaced points n = 30, $|\Omega_i| = 100,\ i = 1,2$.

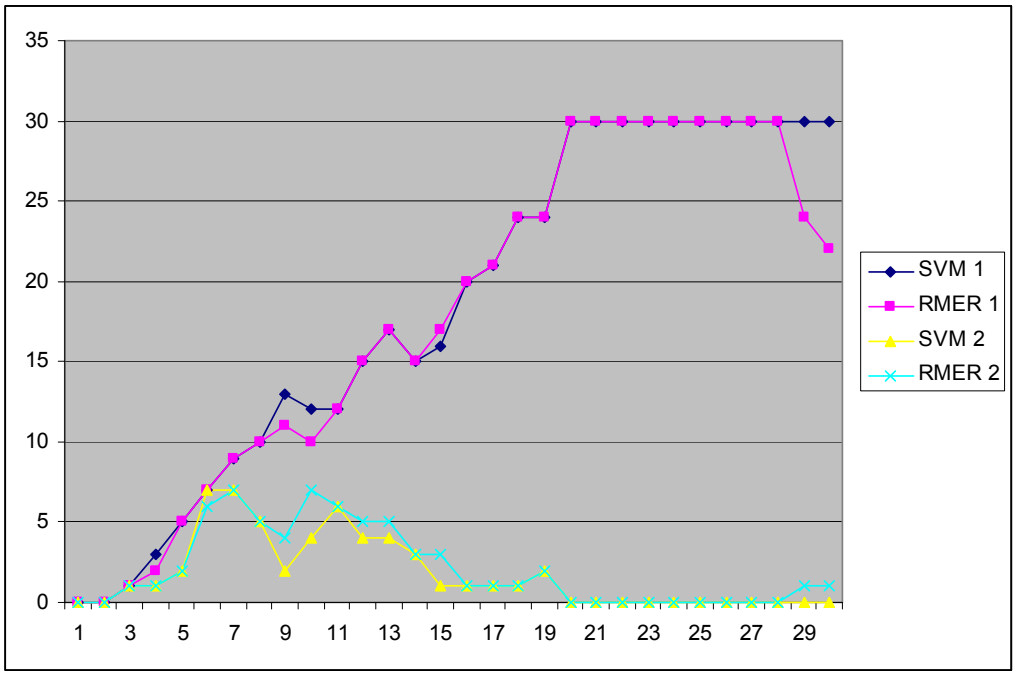


Figure 3. The dependence of classification error on the number of displaced points for each class, n = 30 $|\Omega_1| = 30$, $|\Omega_2| = 200$, SVM 1 - a model of SVM, set Ω_1, SVM 2 - a model of SVM, set Ω_2 , RMER 1 - the relaxed model of minimization of the empirical risk, set Ω_1 , RMER 2 model RMER, set Ω_2.

Bibliography

1. Vapnik V. Statistical Learning Theory. New York: Wiley, 1998.
2. Шлезингер М., Главач В. Десять лекций по статистическому и структурному распознаванию. – К.: Наукова думка, 2004. – 545 с.
3. Thorsten Joachims. Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms. Kluwer, 2002
4. Zhuravlev Yu., Laptin Yu., A.Vinogradov Minimization of empirical risk in linear classifier problem // New Trends in Classification and Data Mining, ITHEA, Sofia, Bulgaria, 2010. – Pages 9-15
5. Журавлев Ю.И., Лаптин Ю.П., Виноградов А.П. Минимизация эмпирического риска и задачи построения линейных классификаторов // Кибернетика и системный анализ. 2011, № 4.- С. 155 – 164.
6. Shor N. Z. Nondifferentiable Optimization and Polynomial Problems. – Amsterdam / Dordrecht / London: Kluwer Academic Publishers, 1998. – 381 p.
7. Yurii I. Zhuravlev, Yuryi Laptin, Alexander Vinogradov, Nikolay Zhurbenko, Aleksey Likhovid. Nonsmooth optimization methods in the problems of constructing a linear classifier // Int Journal Information Models & Analyses (ISSN 1314-6416) 2012 Volume 1 Number 2 pp 103-111.
8. Laptin Yu. P., Likhovid A. P., and Vinogradov A. P. Approaches to Construction of Linear Classifiers in the Case of Many Classes // Pattern Recognition and Image Analysis, Vol. 20, No. 2, 2010, p. 137-145.

Authors' Information

Yurii I. Zhuravlev - *Academician of the RAS, Deputy Director, Dorodnicyn Computing Centre of the RAS, Vavilova 40, 119333 Moscow, Russian Federation; e-mail: zhuravlev@ccas.ru*

Yuryi Laptin - *Senior Researcher, VMGlushkov Institute of Cybernetics of the NASU, Prospekt Academika Glushkova, 40, 03650 Kyiv, Ukraine; e-mail: laptin_yu_p@mail.ru*

Alexander Vinogradov - *Senior Researcher, Dorodnicyn Computing Centre of the RAS, Vavilova 40, 119333 Moscow, Russian Federation; e-mail: vngrccas@mail.ru*

Aleksey Likhovid - *Researcher, VMGlushkov Institute of Cybernetics of the NASU, Prospekt Academika Glushkova, 40, 03650 Kyiv, Ukraine; e-mail: o.lykhovyd @ gmail.com*

ADAPTIVE FUZZY PROBABILISTIC CLUSTERING OF INCOMPLETE DATA

Yevgeniy Bodyanskiy, Alina Shafronenko, Valentyna Volkova

Abstract: *in the paper new recurrent adaptive algorithm for fuzzy clustering of data with missing values is proposed. This algorithm is based on fuzzy probabilistic clustering procedures and self-learning Kohonen's rule using principle "Winner-Takes-More" with Cauchy neighborhood function.*

Using proposed approach it's possible to solve clustering task in on-line mode in situation when the amount of missing values in data is too big.

Keywords*: fuzzy clustering, Kohonen self-organizing network, learning rule, incomplete data with missing values.*

ACM Classification Keywords*: 1.2.6 [Artificial Intelligence]: Learning – Connectionism and neural nets; 1.2.8 [Artificial Intelligence]: Problem Solving, Control Methods, and Search – Control theory; 1.5.1 [Pattern Recognition]: Clustering – Algorithms.*

Introduction

The problem of data sets described by vector-images clustering often occurs in many applications associated with Data Mining, but recently the focus on Fuzzy Clustering [Bezdek, 1981; Hoeppner, 1999; Xu, 2009], when processed vector-image with different levels of probabilities, possibilities or memberships, can belong to more than one class.

However, there are situations when the data sets contain missing values, the information that is lost. In this situation more effective is to use mathematical apparatus of Computational Intelligence [Rutkowski, 2008] and, first of all artificial neural networks [Marwala, 2009], that solve task of restoring the lost observations and modifications of the popular method of fuzzy c-means [Hathaway, 2001], which solve the problem of clustering without recovery of data.

Existing approaches for data processing with missing values [Zagoruyko, 1979; Zagoruyko, 1999], are efficient in cases when the massive of the original observations is given in batch form and does not change during the processing. At the same time, there is a wide class of problems in which the data that arrive to the processing, have the form of sequence that is feed in real time as it occurs in the training of Kohonen self-organizing maps [Kohonen, 1995] or their modifications [Gorshkov, 2009]. In this regard we introduced [Bodyanskiy, 2012] the adaptive neuro-fuzzy Kohonen network to solve the problem of clustering data with gaps based on the strategy of partial distances (PDS FCM). However, in situations where the number of such missing values is too big, the strategy of partial distances may be not effective, and therefore it may be necessary, along with the solution of fuzzy clustering simultaneously estimate the missing observations. In this situation, a more efficient is approach that is based on the optimal expansion strategy (OCS FCM) [Hathaway, 2001]. This work is devoted to the task of on-line data clustering using the optimal expansion strategy, adapted to the case when information is processed in a sequential mode, and its volume is not determined in advance.

Adaptive probabilistic fuzzy clustering of data with missing values based on the optimal expansion strategy

Baseline information for solving the task of clustering in a batch mode is the sample of observations, formed from N n-dimensional feature vectors $X = \{x_1, x_2, ..., x_N\} \subset R^n, x_k \in X, k = 1, 2, ..., N$. The result of clustering is the partition of original data set into m classes $(1 < m < N)$ with some level of membership $U_q(k)$ of k-th feature vector to the q-th cluster $(1 \le q \le m)$. Incoming data previously are centered and standardized by all features, so that all observations belong to the hypercube $[-1,1]^n$. Therefore, the data for clustering form array $\tilde{X} = \{\tilde{x}_1, ..., \tilde{x}_k, ..., \tilde{x}_N\} \subset R^n$, $\tilde{x}_k = (\tilde{x}_{k1}, ..., \tilde{x}_{ki}, ..., \tilde{x}_{kn})^T$, $-1 \le \tilde{x}_{ki} \le 1$, $1 < m < N$, $1 \le q \le m$, $1 \le i \le n$, $1 \le k \le N$ that is, all observations $\tilde{x}_{ki}$ are available for processing.

Introducing the objective function of clustering [Bezdek, 1981]

$$E(U_q(k), w_q) = \sum_{k=1}^{N} \sum_{q=1}^{m} U_q^{\beta}(k) D^2(\tilde{x}_k, w_q)$$

with constraints $\sum_{q=1}^{m} U_q(k) = 1,\ 0 < \sum_{k=1}^{N} U_q(k) < N$ and solving standard nonlinear programming problem, we get the probabilistic fuzzy clustering algorithm [Hoeppner, 1999;Xu, 2009]

$$\begin{cases} U_q^{(\tau+1)}(k) = \dfrac{(D^2(\tilde{x}_k, w_q^{(\tau)}))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (D^2(\tilde{x}_k, w_l^{(\tau)}))^{\frac{1}{1-\beta}}}, \\ w_q^{(\tau+1)} = \dfrac{\sum_{k=1}^{N} (U_q^{(\tau+1)}(k))^{\beta} \tilde{x}_k}{\sum_{k=1}^{N} (U_q^{(\tau+1)}(k))^{\beta}}, \end{cases} \quad (1)$$

where w_q - prototype (centroid) of q-th cluster, $\beta > 1$ - parameter that is called fuzzyfier and defines "vagueness" of boundaries between classes, $D^2(\tilde{x}_k, w_q)$ - the distance between $\tilde{x}_k$ and w_q in adopted metric, $\tau = 0, 1, 2, ...$ - index of epoch of information processing which is organized as a sequence of $w_q^{(0)} \to U_q^{(1)} \to w_q^{(1)} \to U_q^{(2)} \to ...$. The calculation process continues until satisfy the condition

$$\left\| w_q^{(\tau+1)} - w_q^{(\tau)} \right\| \le \varepsilon \ \forall\ 1 \le q \le m,$$

(here ε - defines threshold of accuracy) or until the specified maximum number of epochs Q ($\tau = 0, 1, 2, ..., Q$).

Note also that when $\beta = 2$ and

$$D^2(\tilde{x}_k, w_q) = \left\| \tilde{x}_k - w_q \right\|^2,$$

we get a popular algorithm of Bezdek's fuzzy c-means (FCM)[Bezdek, 1981].

The process of fuzzy clustering can be organized in on-line mode as sequentially processing. At this situation batch algorithm (1) can be rewritten in recurrent form [Bodyanskiy, 2005]

$$\begin{cases} U_q(k+1) = \dfrac{(D^2(\tilde{x}_{k+1}, w_q(k)))^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}(D^2(\tilde{x}_{k+1}, w_l(k)))^{\frac{1}{1-\beta}}}, \\ w_q(k+1) = w_q(k) + \eta(k+1)U_q^{\beta}(k+1)(\tilde{x}_{k+1} - w_q(k)), \end{cases} \quad (2)$$

where $\eta(k+1)$ - learning rate parameter, $U_q^{\beta}(k+1)$ - bell-shaped neighborhood function of neuro-fuzzy Kohonen network (Cauchy function), designed to solve the problems of fuzzy clustering [Gorshkov, 2009; Bodyanskiy, 2012], based on the principle "Winner Takes More» (WTM) [Kohonen, 1995].

In the presence of an unknown number of missing values in vector images $\tilde{x}_k$, that form array $\tilde{X}$, following [Hathaway, 2001], we introduce the sub-arrays:

$X_F = \{\tilde{x}_k \in \tilde{X} \mid \tilde{x}_k - vector\ containing\ all\ components\}$;

$X_P = \{\tilde{x}_{ki}, 1 \le i \le n, 1 \le k \le N \mid values\ \tilde{x}_k,\ available\ in\ \tilde{X}\}$;

$X_G = \{\tilde{x}_{ki} = ?, 1 \le i \le n, 1 \le k \le N \mid values\ \tilde{x}_k, absent\ in\ \tilde{X}\}$.

The optimal expansion strategy consists in the fact that the elements of sub-array X_G are considered as additional variables, which are estimated by minimization of objective function E. Thus, in parallel with clustering (optimization E by $U_q(k)$ and w_q) estimation of missing observations is made (optimization E by $\tilde{x}_{ki} \in X_G$). In this case, the algorithm of fuzzy c-means based on the optimal expansion strategy can be written as the following sequence of steps [Hathaway, 2001]:

1. Setting the initial conditions for the algorithm: $\beta > 0$; $1 < m < N$; $\varepsilon > 0$; $w_q^{(0)}$; $1 \le q \le m$; $\tau = 0,1,2,\ldots,Q$; $X_G^{(0)} = \{-1 \le \hat{x}_{ki}^{(0)} \le 1\}$, where $X_G^{(0)} - N_G (1 \le N_G \le (n-1)N)$ arbitrary initial estimates $\hat{x}_{ki}^{(0)}$ of missing values $\tilde{x}_{ki} \in X_G$;
2. Calculation of membership levels by solving the optimization problem:
$$U_q^{(\tau+1)}(k) = \underset{U_q(k)}{\arg\min} E(U_q(k), w_q^{(\tau)}, X_G^{(\tau)}) = \frac{(D^2(\hat{x}_k^{(\tau)}, w_q^{(\tau)}))^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}(D^2(\hat{x}_k^{(\tau)}, w_l^{(\tau)}))^{\frac{1}{1-\beta}}} = \frac{(\left\|\hat{x}_k^{(\tau)} - w_q^{(\tau)}\right\|^2)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}(\left\|\hat{x}_k^{(\tau)} - w_l^{(\tau)}\right\|^2)^{\frac{1}{1-\beta}}}$$
(here vector $\hat{x}_k^{(\tau)}$ differs from $\tilde{x}_k$ by replacing missing values $\tilde{x}_{ki} \in X_G$ by estimates $\hat{x}_{ki}^{(\tau)}$ that are calculated for the τ-th epoch of data processing);
3. Calculation the centroids of clusters:
$$w_q^{(\tau+1)} = \underset{w_q}{\arg\min} E(U_q^{(\tau+1)}(k), w_q, X_G^{(\tau)}) = \frac{\sum\limits_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta}\hat{x}_k^{(\tau)}}{\sum\limits_{k=1}^{N}(U_q^{(\tau+1)}(k))^{\beta}};$$
4. Checking the stop conditions:
if $\left\|w_q^{(\tau+1)} - w_q^{(\tau)}\right\| < \varepsilon\ \forall\ 1 \le q \le m$ or $\tau = Q$, then the algorithm terminates, otherwise go to step 5;
5. Estimation of missing observations by solving the optimization problem:
$$X_G^{(\tau+1)} = \underset{X_G}{\arg\min} E(U_q^{(\tau+1)}(k), w_q^{(\tau+1)}, X_G)$$

or, equivalently

$$\frac{\partial E(U_q^{(\tau+1)}(k), w_q^{(\tau+1)}, X_G)}{\partial \hat{x}_{ki}} = 0,$$

That leads to

$$\hat{x}_{ki}^{(\tau+1)} = \frac{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k))^{\beta} w_{qi}^{(\tau+1)}}{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k))^{\beta}}.$$

Information processing with this algorithm is organized as a sequence

$$w_q^{(0)} \to U_q^{(1)} \to \hat{x}_{ki}^{(1)} \to w_q^{(1)} \to U_q^{(2)} \to \ldots \to w_q^{(\tau)} \to U_q^{(\tau+1)} \to \hat{x}_{ki}^{(\tau+1)} \to w_q^{(\tau+1)} \to \ldots \to w_q^{(Q)}$$

thus it is possible to organize on-line clustering by type of procedure (2). For this purpose we introduce two time scales: real time $k = 1, 2, \ldots, N, \ldots$, and accelerated computing time $\tau = 0, 1, 2, \ldots, Q$. Here we assume that between two instants of real time k and $k+1$ implemented Q iterations of accelerated time.

Then we can write procedure

$$\begin{cases} U_q^{(\tau+1)}(k+1) = \dfrac{(\left\| \hat{x}_{k+1}^{(\tau)} - w_q(k) \right\|^2)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (\left\| \hat{x}_{k+1}^{(\tau)} - w_l(k) \right\|^2)^{\frac{1}{1-\beta}}}, \\ \hat{x}_{k+1,i}^{(\tau+1)} = \dfrac{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta} w_{qi}(k)}{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta}}, \\ w_q(k+1) = w_q(k) + \eta(k+1)(U_q^{(Q)}(k+1))^{\beta} * \\ \qquad\qquad * (\hat{x}_{k+1}^{(Q)} - w_q(k)), \end{cases} \tag{3}$$

which shows that the memberships and missing observations are calculated in accelerated time, and centroids - in real time by WTM selflearning rule.

Of course centroids can be recalculated in accelerated time too:

$$\begin{cases} U_q^{(\tau+1)}(k+1) = \dfrac{(\left\| \hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1) \right\|^2)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m} (\left\| \hat{x}_{k+1}^{(\tau)} - w_l^{(\tau)}(k) \right\|^2)^{\frac{1}{1-\beta}}}, \\ w_q^{(0)}(k+1) = w_q^{(Q)}(k), \\ w_q^{(\tau+1)}(k+1) = w_q^{(\tau)}(k+1) + \eta(k+1)(U_q^{(\tau+1)}(k+1))^{\beta}(\hat{x}_{k+1}^{(\tau)} - w_q^{(\tau)}(k+1)), \\ \hat{x}_{k+1,i}^{(\tau+1)} = \dfrac{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta} w_{qi}^{(\tau+1)}(k+1)}{\sum_{q=1}^{m} (U_q^{(\tau+1)}(k+1))^{\beta}}, \end{cases} \tag{4}$$

in this case anyway, both in (3) and (4) operation of summation about k is absent, that for large N can involve a lot of memory.

Experiments

Experimental research conducted on two samples of data such as Wine and Iris of UCI repository.

To estimate the quality of the algorithm we used quality criteria partitioning into clusters such as: Partition Coefficient (PC), Classification Entropy (CE), Partition Index (SC), Separation Index (S), Xie and Beni's Index (XB), Dunn's Index (DI) [Xu, 2009].

We also compared the results of our proposed algorithm with other more well-known such as Fuzzy C-means (FCM) clustering algorithm and Gustafson-Kessel clustering algorithm.

The proposed algorithm shown better results than the FCM and Gustafson-Kessel clustering algorithm.

Conclusion

The problem of probabilistic fuzzy adaptive clustering, containing a priori unknown number of gaps, based on the optimal expansion of data strategy is considered. The proposed algorithms are based on the recurrent optimization of a special type of goal functions. Missing observations are replaced by their estimates also obtained in the solution of optimization problem. Centroids of recovered clusters are tuned using a procedure close to the T.Kohonen WTM-rule with the function of the neighborhood (membership), having the Cauchian form.

Acknowledgement

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Bezdek, 1981] J.C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, 1981.

[Bodyanskiy, 2005] Ye. Bodyanskiy. Computational intelligence techniques for data analysis. Lecture Notes in Informatics. Bonn: GI, 2005, V. P-72, P. 15-36.

[Bodyanskiy, 2012] BodyanskiyYe., Shafronenko A., Volkova V. Adaptive clustering of incomplete data using neuro-fuzzy Kohonen network. In “Artificial Intelligence Methods and Techniques for Business and Engineering Applications” – Rzeszow-Sofia: ITHEA, 2012. – P. 287-296.

[Gorshkov, 2009] Ye. Gorshkov, V. Kolodyazhniy, Ye. Bodyanskiy. New recursive learning algorithms for fuzzy Kohonen clustering network. Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronc Systems. (Rapperswil, Switzerland, June 21-24, 2009) Rapperswil, Switzerland, 2009, P. 58-61.

[Hathaway, 2001] R.J. Hathaway, J.C Bezdek. Fuzzy c-means clustering of incomplete data. IEEE Trans. on Systems, Man, and Cybernetics, №5, 31, 2001, P. 735-744.

[Hoeppner, 1999] F Hoeppner, F. Klawonn, R. Kruse, T. Runker. Fuzzy Clustering Analysis: Methods for Classification, Data Analysis and Image Recognition. Chichester, John Wiley &Sons, 1999.

[Kohonen, 1995] T. Kohonen. Self-Organizing Maps. Berlin: Springer-Verlag, 1995.

[Krishnapuram, 1993] R. Krishnapuram, J.M. Keller. A possibilistic approach to clustering. Fuzzy Systems, 1993, 1, №2, P.98-110.

[Marwala, 2009] T Marwala. Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques. Hershey-New York, Information Science Reference, 2009.

[Rutkowski, 2008] L.Rutkowski. Computational Intelligence.Methods and Techniques. Berlin-Heidelberg: Springer-Verlag, 2008.

[Xu, 2009] R. Xu, D.C. Wunsch. Clustering. Hoboken, N.J. John Wiley & Sons, Inc., 2009.

[Zagoruyko, 1979] N.G. Zagoruyko. Empirical predictions. Novosibirsk, Nauka, 1979 (in Russian).

[Zagoruyko, 1999] N.G. Zagoruyko. Applied Data Analysis and Knowledge. Novosibirsk, 1999 (in Russian).

Authors' Information

Yevgeniy Bodyanskiy – Professor, Dr. – Ing. habil., Scientific Head of Control Systems Research Laboratory, Kharkiv National University of Radio Electronics, 14 Lenin Ave., Office 511, 61166 Kharkiv, Ukraine; e-mail: bodya@kture.kharkov.ua

Major Fields of Scientific Research: Artificial neural networks, Fuzzy systems, Hybrid systems of computational intelligence

Alina Shafronenko – Ph.D student in Artificial Intelligence dept., Kharkiv National University of Radioelectronics Lenin Ave., 14, Kharkiv, 61166, Ukraine; e-mail: alinashafronenko@gmail.com

Major Fields of Scientific Research: neural networks, neural network processing of data with gaps, fuzzy clustering, clustering of data

Valentyna Volkova *- Candidate of Technical Science (Ph.D.), Senior lecturer in Artificial Intelligence dept., Kharkiv National University of Radioelectronics Lenin Ave., 14, Kharkiv, 61166, Ukraine; e-mail: volkova@kture.kharkov.ua*

Major Fields of Scientific Research: neural networks, fuzzy clustering, clustering of data

CROP CLASSIFICATION IN UKRAINE USING SATELLITE OPTICAL AND SAR IMAGES

Nataliia Kussul, Sergii Skakun, Andrii Shelestov, Oleksii Kravchenko, Olga Kussul

***Abstract**: This paper presents first results of the use of optical and synthetic-aperture radar (SAR) satellite images to crop classification in Ukraine. The study aims at optimizing SAR parameters to provide timely and economically efficient crop maps/area estimates for Ukrainian landscape. Integration of EO-1 and RADARSAT-2 is done for crop classification using support vector machine (SVM). Classification is carried out per-field. The results on using SAR data for summer crops classification look very promising enabling classification of major summer crops with 10-15% of commission and omission errors under a typical Ukrainian landscape.*

***Keywords**: crop classification, SVM, satellite images, SAR multipolarization.*

***ACM Classification Keywords**: I.4.8 [Image Processing and Computer Vision] Scene Analysis - Sensor Fusion.*

Introduction

Space Research Institute of NAS Ukraine and SSA Ukraine (SRI, Ukraine) is actively involved in the Joint Experiment for Crop Assessment and Monitoring (JECAM) activities of the Group on Earth Observations (GEO), and established two JECAM test sites in Ukraine in 2011 [Kussul et al, 2012]. One of the main problems being solved is crop identification and crop area estimation with the use of satellite optical and synthetic-aperture radar (SAR) images.

Recently, many studies have focused on the estimation of crop acreage and proportions at global scale using time-series of moderate resolution remote sensing images such as MODIS and SPOT-VEGETATION [Verbeiren et al, 2008; Fritz et al, 2008; Pan et al, 2012; Wu and Li, 2012; Vintrou et al, 2012]. The use of synthetic-aperture radar (SAR) images in combination with optical ones for crop acreage estimation and crop monitoring is discussed in [McNairn et al, 2009a,b; Leichtle et al, 2012; Blaes et al, 2005; Skriver et al, 2011; Hoekman et al, 2011]. As to classification algorithms, decision tree (DT) [McNairn et al, 2009a,b; Vintrou et al, 2012; Boryan et al, 2008], artificial neural networks (ANN) [Verbeiren et al, 2008; Fritz et al, 2008; Gallego et al, 2012; Skakun et al, 2007; Kussul et al, 2012], support vector machine (SVM) [Gallego et al, 2012; Kussul et al, 2012; Pal and Foody, 2012], and maximum likelihood (ML) [Pan et al, 2012; Wu and Liu, 2012; Skriver et al, 2011] have been recently the most widely used. In general, overall classification accuracy reported in the literature is 80% to 90% depending on the available remote sensing images, complexity of landscape and extent of the region.

In 2010, SRI completed an EC-JRC contract “Crop area estimation with satellite images in Ukraine” within the MARS and GEOLAND2 projects which showed particular difficulties in discriminating summer crops using satellite optical images [Gallego et al, 2012; Kussul et al, 2012]. This was due to cloud cover, not optimal dates of satellite image acquisition and inherent limits of optical data. Therefore, feasibility and use of satellite synthetic-aperture radar (SAR) multi-polarized images need to be exploited.

This problem is addressed within the SRI project of the SOAR-JECAM program to acquire quad-polarized RADARSAT-2 images. The project aims at optimizing SAR parameters to provide timely and economically efficient crop maps/area estimates for Ukrainian landscape [Kogan et al, 2013; Shelestov et al, 2013; Kussul et al, 2011; Kussul et al, 2010]. This paper discusses the results of integration of SAR and optical satellite images for crop classification in Ukraine.

Study area and data description

For exploring capabilities of satellite SAR imagery to discriminate summer crops in Ukraine, a test site of Vasylkiv county in Kyivska oblast was selected (Fig. 1). The test site incorporates all major summer crops (maize, soy beans, sunflower, sugar beet), and has total area of about 1,200 sq. km.

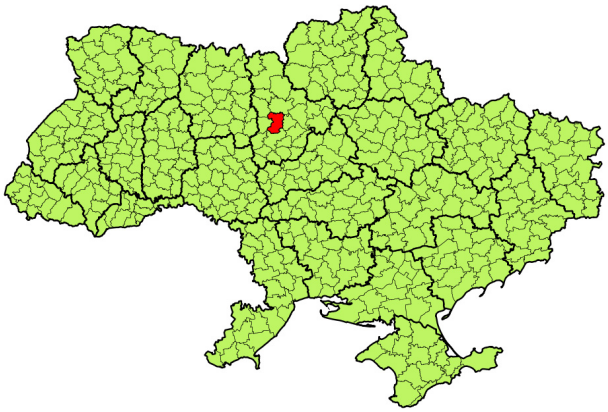
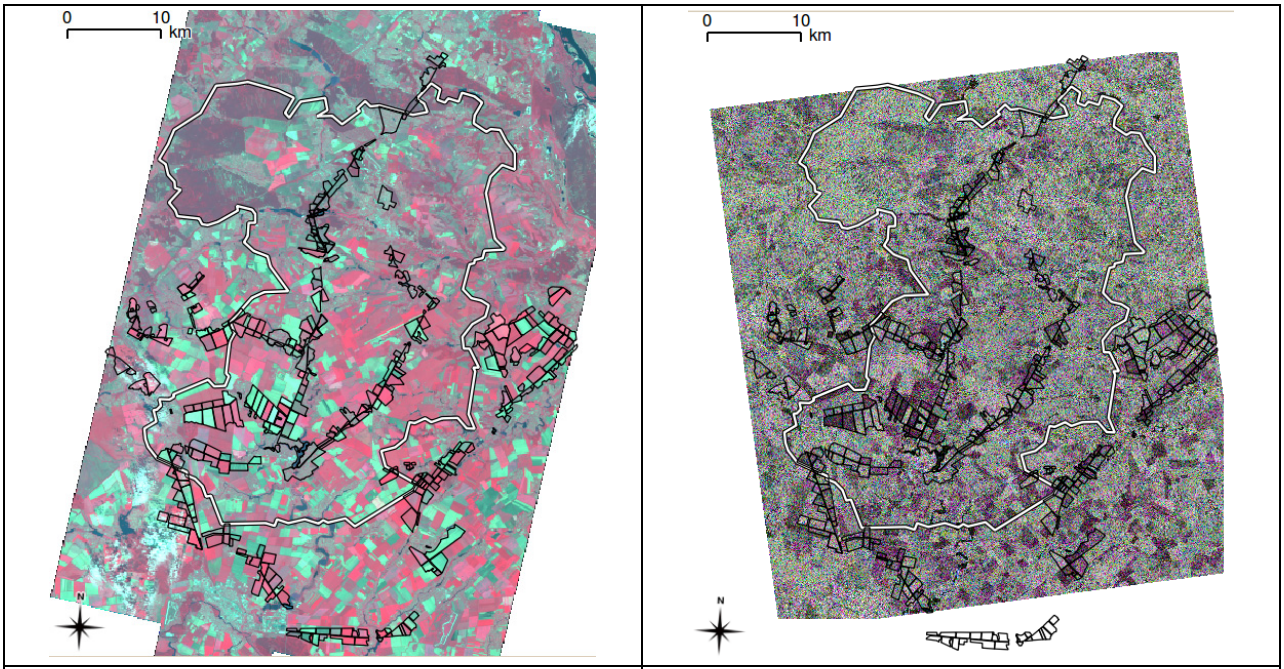


Figure 1. The Vasylkiv county test site shown in red

The first RADARSAT-2 images over one of the test-site were acquired on 27 and 30 July 2012. Optical images from EO-1 satellite were acquired on 28 July 2012 as well. Ground observations to support satellite images were carried out on 4 August 2012 to collect in situ measurements (Fig. 2). In total, information on 271 fields was collected. The following crop types were present (Table 1): maize, soy beans, sunflower, sugar beet, harvested winter and spring crops, and minor crops like buckwheat.



(a) (b)

Figure 2. Boundaries of observed crop fields outlined in black over EO-1 VNIR image (a) and RADARSAT 2 quad-pol images (b). The test site area is shown in white. EO-1 Data are courtesy of the NASA Earth Observing One (EO-1) mission operated by the Goddard Space Flight Center, 2012. RADARSAT-2 Data and Products © MacDonald, Dettwiler and Associates Ltd.(2012) - All Rights Reserved. RADARSAT is an official trademark of the Canadian Space Agency.

Table 1. Ground survey statistics

Crop	Fields observed, %	Total area (ha), %
Maize	63 (23.2%)	6264 (29.5 %)
Soy beans	46 (17.0%)	3553 (16.7 %)
Sunflower	19 (7.0%)	1722 (8.1 %)
Sugar beet	12 (4.4%)	1130 (5.3 %)
Winter and spring crops (already harvested)	79 (29.2 %)	6184 (29.1 %)
Non agricultural land and minor crops	52 (19.2 %)	2369 (11.2 %)
Total	271 (100 %)	21222 (100 %)

Classification method

SAR images were classified using a per-field classification approach. Many previous studies have shown that efficiency of SAR classification can be improved using per-field approach due to the presence of speckle at the pixel level [Blaes et al, 2005]. For each field, a median value of backscatter coefficient was estimated for each polarization (VV, VH, HV, and HH), and was used as an input to classification algorithm. Classification was done using the SVM classifier.

Support vector machine (SVM) became popular for solving problems in classification, regression, and novelty detection [Bishop, 2006]. An important property of support vector machines is that the determination of the model parameters corresponds to a convex optimization problem, and so any local solution is also a global optimum. The SVM is a decision machine and, unlike ANNs, does not provide posterior probabilities. Also, processing new datasets could be resource consuming comparing to the ANNs.

Classification results

Collected data was used for preliminary analysis of the discriminating power of optical, SAR and a combination of them both by visual interpretation and supervised classification. In contrast to optical images visual interpretation of SAR images allows distinguishing maize, soy beans and combined sunflower & sugar beet crops due to different canopy architecture and different scattering processes. Sugar beet and sunflower could not be discriminated.

Numerical analysis was performed using a Support Vector Machine (SVM) classifier. Classification accuracies, commission and omission errors were estimated using a five-fold cross-validation procedure. Special care has been taken to prevent over-fitting of cross-validation procedure due to spatial correlation in collected data. Three different data sets were examined: combined EO-1 and RADARSAT-2, EO-1 only and RADARSAT-2 only data (Table 2).

Total accuracies appear similar in all datasets because they are heavily influenced by winter and spring crop class that is classified equally well by optical and SAR data.

The difference between datasets lies in per crop classification errors. All summer crops are better classified using SAR data rather then optical data. The most profound effect is observed on soy beans that is the second major summer crop after maize in the given area. Using SAR data instead of optical allows decreasing omission error for soybeans from 34% to 13% while maintaining similar level of commission error.

Combined dataset shows gradual decrease of errors for most crops 5% to 10%. Sunflower + sugar beet class is

the most beneficial as combined data allows decreasing classification errors in 1.5-2 times in comparison with optical or SAR data alone. This result is explained by complementary roles of SAR and optical data for discriminating sunflower at flowering phenological stages.

Table 2. Total accuracy, commission (CE) and omission (OE) classification errors

Crop	**EO1 + R2**		**EO1**		**R2**	
Total accuracy	**91.4 %**		**84.8 %**		**88.9 %**	
	CE, %	**OE, %**	**CE, %**	**OE, %**	**CE, %**	**OE, %**
Maize	11.8 %	4.8 %	17.6 %	11.1 %	13.6 %	9.5 %
Soy beans	6.1 %	18.4 %	7.4 %	34.2 %	10.8 %	13.2 %
Sunflower + sugar beet	20.0 %	23.1 %	39.3 %	34.6 %	28.0 %	30.8 %
Winter + spring	2.8 %	1.4 %	6.7 %	1.4%	2.9 %	4.2 %

Conclusions

The first results on using SAR data for summer crops classification look very promising enabling classification of major summer crops with 10-15% of commission and omission errors under a typical Ukrainian landscape. Further research is required to ensure robustness of proposed approach on larger areas. Additional efforts should be put on investigation of possibility to substitute quad-polarization SAR data with cheaper wide swath dual-polarization data available from RADARSAT-2 and the upcoming Sentinel-1 satellites.

Bibliography

[Bishop, 2006] C. Bishop. Pattern Recognition and Machine Learning, Springer, New York, USA, 2006.

[Blaes et al, 2005] X. Blaes, L. Vanhalleb, and P. Defourny. Efficiency of crop identification based on optical and SAR image time series. Remote Sens. of Env., 96, pp. 352–365, 2005.

[Boryan et al, 2008] C. Boryan, M. Craig, and M. Lindsay. Deriving Essential Dates of AWiFS and MODIS Data for the Identification of Corn and Soybean Fields in the U.S. Heartland. In: Pecora 17 – The Future of Land Imaging. Going Operational, Denver, USA, 2008.

[Fritz et al, 2008] S. Fritz, M. Massart, I. Savin, J. Gallego, and F. Rembold. The use of MODIS data to derive acreage estimations for larger fields: A case study in the south-western Rostov region of Russia. Int. J. of Appl. Earth Observ. Geoinf., vol. 10, no. 4, pp. 453–466, 2008.

[Gallego et al, 2012] J. Gallego, A. Kravchenko, N. Kussul, S. Skakun, A. Shelestov, and Y. Grypych,. Efficiency Assessment of Different Approaches to Crop Classification Based on Satellite and Ground Observations. J. of Automation and Inf. Sci., vol. 44, no. 5, pp. 67–80, 2012.

[Hoekman et al, 2011] D.H. Hoekman, M.A.M. Vissers, T.N. Tran. Unsupervised Full-Polarimetric SAR Data Segmentation as a Tool for Classification of Agricultural Areas,. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 4, no. 2, pp. 402–411, 2011.

[Kogan et al, 2013] F. Kogan, N. Kussul, T. Adamenko, S. Skakun, O. Kravchenko, O. Kryvobok, A. Shelestov, A. Kolotii, O. Kussul and A. Lavrenyuk. Winter wheat yield forecasting in Ukraine based on Earth observation, meteorological data and biophysical models. Int. J. Appl. Earth Observ. Geoinf., vol. 23, pp. 192–203, 2013.

[Kussul et al, 2010] N. Kussul, A. Shelestov, S. Skakun, O. Kravchenko. High performance Intelligent Computations for Environmental and Disaster Monitoring. In: Intelligent Data Processing in Global Monitoring for Environment and Security (Krassimir Markov, Vitalii Velychko editors), I T H E A, Sofia, pp. 64-92, 2010.

[Kussul et al, 2011] N. Kussul, S. Skakun, O. Kravchenko. Environmental Risk Assessment Using Geospatial Data and Intelligent Methods. Int. J. "Information Technologies & Knowledge", Vol.5, Number 2, pp. 129-140, 2011.

[Kussul et al, 2012] N. Kussul, A. Shelestov, S. Skakun, O. Kravchenko, B. Moloshnii. Crop state and area estimation in Ukraine based on remote and in-situ observations. Int. J. on Information Models and Analyses, vol. 1, no. 3, pp. 251-259, 2012.

[Kussul et al, 2012] N. Kussul, S. Skakun, A. Shelestov, O. Kravchenko, J. F. Gallego, and O. Kussul. Crop area estimation in Ukraine using satellite data within the MARS project. In: 2012 IEEE Int. Geoscience and Remote Sensing Symposium,

22-27 July 2012, Munich, Germany, pp. 3756–3759, 2012.
[Leichtle et al, 2012] T. Leichtle, A. Schmitt, A. Roth, and M. Schardt. On the capability of different SAR polarization combinations for agricultural monitoring. In: 2012 IEEE Int. Geoscience and Remote Sensing Symposium, 22-27 July 2012, Munich, Germany, pp. 3752–3755, 2012.
[McNairn et al, 2009a] H. McNairn, J. Shang, C. Champagne, and X. Jiao. TerraSAR-X and RADARSAT-2 for crop classification and acreage estimation. In: 2009 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2009, vol.2, pp. II-898–II-901, 12-17 July 2009.
[McNairn et al, 2009b] H. McNairn, C. Champagne, J. Shang, S. Holmstrom, and G. Reichert. Integration of optical and Synthetic Aperture Radar (SAR) imagery for delivering operational annual crop inventories. ISPRS J. of Photogramm. and Remote Sens., 64(5), pp. 434–449, 2009.
[Pal and Foody, 2012] M. Pal, and G.M. Foody. Evaluation of SVM, RVM and SMLR for Accurate Image Classification With Limited Ground Data. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 5, no. 5, pp. 1344–1355, 2012.
[Pan et al, 2012] Y. Pan, L. Li, J. Zhang, S. Liang, X. Zhu, and D. Sulla-Menashe. Winter wheat area estimation from MODIS-EVI time series data using the Crop Proportion Phenology Index. Remote Sens. of Env., vol. 119, pp. 232–242, 2012.
[Shelestov et al, 2013] A.Yu. Shelestov, A.N. Kravchenko, S.V Skakun., S.V. Voloshin, and N.N. Kussul. Geospatial information system for agricultural monitoring. Cybern. Syst. Anal., vol. 49, no. 1, pp 124–132, 2013.
[Skakun et al, 2007] S. Skakun, E. Nasuro, A. Lavrenyuk, and O. Kussul. Analysis of Applicability of Neural Networks for Classification of Satellite Data. J. of Automation and Inf. Sci., vol. 39, no. 3, pp. 37–50, 2007.
[Skriver et al, 2011] H. Skriver, F. Mattia, G. Satalino, A. Balenzano, V. R. Pauwels, N. E. Verhoest, and M. Davidson. Crop classification using short-revisit multitemporal SAR data. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 4, no. 2, pp. 423-431, 2011.
[Verbeiren et al, 2008] S. Verbeiren, H. Eerens, I. Piccard, I. Bauwens, and J. Van Orshoven. Sub-pixel classification of SPOT-VEGETATION time series for the assessment of regional crop areas in Belgium. Int. J. of Appl. Earth Observ. Geoinf., vol. 10, no. 4, pp. 486–497, 2008.
[Vintrou et al, 2012] E. Vintrou, A. Desbrosse, A. Bégué, S. Traoré, C. Baron, and D. Lo Seen. Crop area mapping in West Africa using landscape stratification of MODIS time series and comparison with existing global land products. Int. J. of Appl. Earth Observ. Geoinf., vol. 14, no. 1, pp. 83–93, 2012.
[Wu and Liu, 2012] B. Wu, and Q. Li. Crop planting and type proportion method for crop acreage estimation of complex agricultural landscapes. Int. J. of Appl. Earth Observ. Geoinf., vol. 16, pp. 101–112, 2012.

Acknowledgement

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and ADUIS (www.aduis.com.ua).

Authors' Information

Nataliia Kussul – *Deputy Director, Head of Department of Space Information Technologies and Systems, Space Research Institute NASU-SSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine; e-mail: inform@ikd.kiev.ua*
Major Fields of Scientific Research: Grid computing, Sensor Web, Earth observation, satellite data processing, risk analysis.

Sergii Skakun *– Head of Laboratory for Satellite Monitoring, Space Research Institute NASU-SSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine; e-mail: serhiy.skakun@ikd.kiev.ua*
Major Fields of Scientific Research: Grid computing, Sensor Web, Earth observation, satellite data processing, risk analysis.

Andrii Shelestov – *Head of Department of Software Engineering, National University of Life and Environmental Sciences of Ukraine, Heroyiv Oborony st., 15, Kyiv-03041, Ukraine; e-mail: andrii.shelestov@gmail.com*
Major Fields of Scientific Research: Grid computing, distributes systems, system architecture design, Earth observation, satellite data processing.

Oleksii Kravchenko *– Senior Scientist, Space Research Institute NASU-SSAU, Glushkov Prospekt 40, build. 4/1, Kyiv 03680, Ukraine; e-mail: oleksiy.kravchenko@gmail.com*
Major Fields of Scientific Research: Earth observation, satellite data processing, machine learning.

Olga Kussul *– Assistant Professor, Physics & Technology Institute of the National Technical University of Ukraine "Kiev Polytechnic Institute", 37 Prospect Peremogy, Kyiv 03056, Ukraine; e-mail: olgakussul@gmail.com*
Major Fields of Scientific Research: trust management in distributed and heterogeneous systems, satellite data processing.

USE OF INFORMATION VALUE IN AVO-POLYNOMIAL METHOD TRAINING

Alexander Dokukin

Abstract: A new approach for constructing an algorithmic basis for polynomial based recognition method is considered. On the contrary to the previously used technical approach based on polynomial power minimization the new one deals with information value of the items. The approaches are compared in AVO-polynomial framework over a same set of recognition tasks.

Keywords: *estimates calculating algorithm, information value, algebraic approach, pattern recognition.*

ACM Classification Keywords*: I.5 Pattern Recognition — I.5.0 General.*

Introduction

The principles of achieving basic set of algorithms becomes important when using algebraic approach for solving actual recognition tasks. It can be some technical approach, a polynomial power minimization for example, which underlies the particular AVO-polynomial recognition method [Dokukin, 2009]. Or it can be recognition quality of separate algorithms and so on.

In this work we offer a new approach based on information value optimization and then compare it to the well studied one. Moreover we try several information value estimates for that purpose including some heuristics as well as statistical and entropic ones.

The first part of the article will be devoted to recalling some necessary definitions and statements. That is standard form of recognition task, estimates calculating algorithm (ECA) and so on.

Then we will describe ECA height minimization task and method from which we will derive information value definition for ECA. In that part an AVO-polynomial method will be described also which correspond to a polynomial over minimal height ECAs.

Finally a testing framework will be described based on AVO-polynomial modifications and the two approaches will be compared by solving actual recognition tasks.

Definitions

The standard recognition task [Zhuravlev, 1977a] is stated as follows. Let us have a training sample $\{S_1, \ldots, S_{m+q}\}$, described by vectors of some nature $S_i = (a_{i1}, \ldots, a_{in})$. The sample is split into l classes $K_1, \ldots, K_l$ that can overlap in general case. The training sample classification that is vectors $\boldsymbol{\alpha_i} = (\alpha_{i1}, \ldots, \alpha_{il})$ is known. Here α_{ij} is the value of "$S_i \in K_j$" predicate. It is required to construct an algorithm A that can calculate classification of a new object S.

Estimates calculating algorithm (ECA or AVO in Russian) refers to a general family of recognition algorithms described by Yu. I. Zhuravlev in 1970s [Zhuravlev, 1977b]. In AVO-polynomial a small subset of the family is used, which corresponds to a case of a single supporting set and equal weights of training objects.

First, it is supposed that all the features are real numbers. A vector $(\varepsilon_1, \ldots, \varepsilon_n)$ of nonnegative ε-thresholds is used to define a proximity function of two objects $B(S_u, S_v)$:

$$B(S_u, S_v) = \begin{cases} 1, \ |a_{ui} - a_{vi}| \leq \varepsilon_i, \ i = 1, \ldots, n, \\ 0, \ \text{otherwise.} \end{cases}$$

Second, object estimates for different classes are introduced, which in the current case are transformed to the following:

$$\Gamma_j(S) = \sum_{S_i \in \tilde{K}_j} B_\omega(S, S_i) \ .$$

Here $\Gamma_j(S)$ is an estimate of an object's S belonging to a K_j, $\tilde{K}_j = \{S_1, \ldots, S_m\} \cap K_j$ and $C\tilde{K}_j = \{S_1, \ldots, S_m\} \setminus K_j$.

Finally, a decision rule is applied converting real-value class estimates to a final decision. In the described case an object is assigned to a class of maximal estimate.

The AVO-polynomial [Dokukin, 2009] method represents a polynomial over ECAs as it is supposed by its name. Algebraic operations are applied to the class estimates before the decision rule. That is the essence of the algebraic approach to recognition [Zhuravlev, 1977a].

This particular polynomial is build according to the following procedure. A set of objects $\{S^1, \ldots, S^q\}$ is extracted from the training sample which will be named a reference sample. As opposed to a testing one, the reference sample is used in training too, but has a different role. Each member S^t of the reference set in combination with all the remaining training objects is used for constructing one item of the polynomial B_t. The following formula describes the polynomial

$$B(S) = \sum_{S^t \in \{S^1, \ldots, S^q\}} D(S) B_t(S) \ .$$

Here B_t is an ECA achieved through training and $D(S)$ is another ECA multiplier penalizing remoteness from S_t, which make the construction a second degree polynomial.

To achieve the B_t an auxiliary task is considered. For each remaining training object $S_i = (a_{i1}, \ldots, a_{in})$ and the $S^t = (b_{t1}, \ldots, b_{tn})$ a new one is constructed $S = (|a_{i1} - bt1|, \ldots, |a_{in} - b_{tn}|)$. The object is assigned to a class 1 if both ancestors belong to a same class and to a class 0 otherwise. Then the optimal hyper-parallelepiped R (rectangle for short) is searched maximizing the difference between class 1 objects number and class 0 ones (ECA height).

Information value

An introduction of the auxiliary task allows to consider the ECA items from a different point of view. Indeed, each one of them is assigned 4 values corresponding to the number of objects in respect to the two predecates: "an object belongs to the class 1" and "an object belongs to the rectangle". Thus, corresponding ECAs can be characterized by an information value similar to the logical regularities methods [Ryazanov, 2007].

A set of different information values estimates will be used further, which is taken from [Vorontsov, 2007], among them some heuristics (norm, ratio, weighted difference), the statistical one, and the IGain.

Let's denote by P and N a number of objects of classes 1 and 0 correspondingly. A number of them belonging to a rectangle will be denoted by p and n. Let's cite the strict formulae of the target functionals used further.

1. Norm.
$$f(R) = \frac{p}{n+p} \ .$$
2. Ratio.
$$f(R) = \frac{p}{n+1} \ .$$
3. Weighted difference.
$$f(R) = \frac{p}{P} - \frac{n}{N} \ .$$

4. Statistical.

$$f(R) = -\ln\left(\frac{C_P^p C_N^n}{C_{P+N}^{p+n}}\right) .$$

5. IGain.

$$f(R) = \hat{H}(P,N) - \hat{H}_\phi(P,N,p,n) ;$$

where

$$\hat{H}(P,N) = H(\frac{P}{P+N}, \frac{N}{P+N}) , \; H(q_0,q_1) = -q_0\log_2(q_0) - q_1\log_2(q_1) ,$$

$$\hat{H}_\phi(P,N,p,n) = \frac{p+n}{P+N}\hat{H}(P,N) + \frac{P+N-p-n}{P+N}\hat{H}(P-p,N-n) .$$

In this notation the ECA height functional is defined as $p-n$.

Comparison

The comparison of different functionals for ECA optimization was performed using AVO-polynomial modifications and real UCI-repository problems. On the first stage a fixed number of ECA items were constructed according to the functional chosen. After that a polynomial was constructed in a same way for each test. Then, AVO-polynomial recognition quality was tested. In each case the quality were averaged over 100 different random partitionings to the training and testing subsets. The results are shown in the following table.

Table 1: Comparison results

Task	Height	Information Value				
		Norm	Ratio	Weighted Difference	Statistical	IGain
Hepatitis	71.9	69.8	**80.2**	71.1	75.4	68.3
Credit	86.5	73.9	73.5	**87.7**	**87.8**	69.7
Echocardiogram	66.9	58.8	61.6	67.5	**71**	61.9
Glass	80.6	79.1	81.9	**82.3**	75.6	70.7
Wine	95.7	95.3	**97.1**	**97.3**	96.2	90

Conclusion

1. In major part of tasks the height optimization performed better than average of other functionals even if leave only the three leading ones.

2. Again for a major of tasks the height performace was comparable to a leading one yielding up to 1.5% percent.

3. In some cases an alternative approach allowed improving the result up to 7%.

Thus, it is effectual introducing a new parameter to the AVO-polynomial training scheme corresponding to an optimization functional.

Acknowledgements

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

Zhuravlev Yu. I. Correct algebras over sets of incorrect (heuristic) algorithms I (in Russian) // Cybernetics. — 1977. — No. 4. — Pp. 14–21.

Zhuravlev Yu. I. Correct algebras over sets of incorrect (heuristic) algorithms II (in Russian) // Cybernetics. — 1977. — No. 6. — Pp. 21–27.

Ryazanov V. V. Logical regularities in pattern recognition (parametric approach) // Computational Mathematics and Mathematical Physics. — 2007. — Vol. 47, No. 10. — Pp. 1720–1735.

Vorontsov K. V. Lections on classification algorithms (in Russian). — 2007. — URL: http://www.ccas.ru/voron/download/LogicAlgs.pdf.

Dokukin A. A. “AVO-Polynom” Recognition Algorithm // Supplement to International Journal “Information Technologies and Knowledge”. — 2009. — Vol. 3, Book 8 “Classification, Forecasting, Data Mining”. — Pp. 65–68.

Authors' Information

Alexander Dokukin *— Computing Centre of Russian Academy of Sciences, researcher, 40 Vavilova St., Moscow, Russia, 119333; e-mail: dalex@ccas.ru.*
Major Fields of Scientific Research: Algebraic Approach to Pattern Recognition.

SHORT GRAPH-SCHEME OF A SUCCESSFUL SYSTEM IDEA

Nikolay Kosovskiy

Abstract*: Notions of a successful system and an ecologically acceptable successful system are proposed. These notions are convenient for formalization of a received inventive problem solution idea by means of TRIZ. This name was proposed by G.S. Altshuller [1] as abbreviation of the Theory of Inventive Problem Solution. Such a formalization is useful for training a creator of a successful system or an ecologically acceptable successful system for the intensification of his creative efforts.*

Keywords*: inventive graph-scheme idea,system, sub-system, ,graph, algorithmic heuristic, TRIZ.*

ACM Classification Keywords*: H.1.1 MODELS AND PRINCIPLES Systems and Information Theory – General systems theory; I.6.1 SIMULATION AND MODELING Simulation Theory – Systems theory*

Introduction

At our days a heuristic technology named in Russian TRIZ (Theory of Inventive Problem Solution) is widely spread. A mathematical model of an invention is introduced in this paper. Below the term “system” is used for a description of inventive scheme idea aimed to the solution of a practical problem. For a formal description of the result of a practical invention problem solution two notions are proposed: a successful system and an ecologically acceptable successful system [2].

Preliminary considerations

A successful system is a pair (invention scheme; numerical characteristic of this scheme outcome). Every successful system may be represented by a graph of relations between its elements names. Every element of a successful system may be also regarded as a successful system. The term “successful sub-system” will be used for such an element.

An ecologically acceptable successful system is a successful system with sufficiently precisely measurable parameters and their bounds which are ecologically admissible for the nature.

A set of a system elements for which a numerical result of its work as well as its productivity may be pointed out is called a successful sub-system.

A graph-scheme of system is a graph with nodes and edges marked by means of short names. A successful scheme of system requires also the presence of sufficiently precisely measurable system output.

It is useful to point out actions (named fields in TRIZ) upon some system elements (named objects). Such a graph-scheme of a system is called in TRIZ “OBACT” (as abbreviation of the words OBject and ACTion). In Russian the term “VEPOL” is used instead of OBACT. So, a graph-scheme of a successful system uses two sorts of elements: objects and actions.

Consider, for example, the **problem of saving a skier** who breaks thin ice not far from his fellows. The fellows are on sufficiently stable ice. The skier in the water may be regarded as an object. The pulling by a stick may be

considered as an action. Unfortunately such an action directly does not allow to pull the skier in wet clothes from the polynia. The edge of stable ice and the weight of the skier prevent this.

The problem may be solved by means of an additional object which is the skier who put his hand with a stick and HIS FOOT WITH A SKI on the stable ice. In such a case the foot with a ski allows the skier to get out on the stable ice.

Here an action directly upon the object (which does not solve the problem) is replaced by an action upon an additional object. So we replace one object by an additional one in OBACT.

While solving an inventive problem it is important to detect a key (the most important) conflict (contradiction). Criteria of primari importance of a necessity may be used.

It is needed to mark that the design of OBACT is a creative one. The main essence of a successful system must be represented in OBACT. Nevertheless the unnecessary detailing of some its sub-systems must be avoid. It is convenient to consider such a sub-system as an object or as an action.

OBACT may be regarded as a set of atomic formulas i.e. predicates of objects which are the ends of oriented edges beginning in an action. In such a case every predicate and every object is presented in OBACT only once.

Somebody can see in OBACT some similarity with a mathematical notion of category using objects and morphisms. Another ones can consider OBACT as an ontological approach to the knowledge representation.

Algorithmic heuristic

As a sub-system and an above-system is considered simultaneously with a system, so it is useful to consider a sub-object and an above-object simultaneously with an object as well as a sub-action and an above- action simultaneously with an action.

The edges of scheme of system graph may be oriented and non-oriented, soft and strong, initial and final, insufficient and excess, etc.

Comparison of two successful systems with each other may be done first of all on the base of the efficiency (success) of their outputs.

In the opposite to the system of mathematical problem solving offered by G. Polya [3, 4] and aimed to stimulation of appearance of ideas useful for the solving of a problem, TRIZ uses trends in the technical system development. Such a trend must be formulated as a low, for example, the low of system transformation to a micro-system.

The described here approach may be called an algorithmic heuristic. But such a name is rather conflicting. If an algorithm exists then there is no place for heuristic. If it is only heuristic then we can't say about an algorithm. More precise name for such an approach would be “an algorithmically organized set of heuristics on the different levels of successful system creating and, first of all, on the level of its idea description” .

Namely, I speak about some ways of human creative ability intensification aimed to the improvement of a successful system. The algorithmic heuristic helps to receive an answer for the question: ”How to guess hidden but sometimes evident solution of a problem?”. But such an “evidence” often becomes clear only after the solving of a problem. The decreasing of heuristic exhaustive search may be done, in particular, by means of heuristics structuring, set and making exhaustive search on the deeper level. I use also some modification of brain storming in student seminars.

Briefly the described algorithmic approach may be characterized as one oriented to the nonstandard solving of a practical problem for the improvement of successful system by means of exhaustive search of heuristics organized as a hierarchy structure. It may be represented as a heuristic technology.

More detailed exposition of educational technology for innovative solution of an inventive problem is presented in [2].

Solving of mathematical problems

The notion of successful system may be extended to a mathematical proof. Detailed algorithm as well as every other result received while solving a nonstandard mathematical problem beginning from the school level and up to the university one. The term “nonstandard mathematical problem” is used for such a problem which has no formal algorithm of its decision. For example, the main table from the book [3] has a two levels. The higher level contains only four useful questions:

- Are you ready to understand more exactly the problem statement?
- Are you ready to make a decision plan?
- Are you ready to implement the formulated decision plan?
- Are you ready to investigate the received solution?

Below the sequence of useful recommendations helping to answer the second question is formulated. This recommendations were formulated by me according to similar ones from different authors. The maximal degree of recommendation and the maximal degree of creative application of recommendation are very importante while using this sequence. The offered recommendations are the most useful for the solving of nonstandard problems. The sequence of such recommendations may be named an algorithmic heuristic one.

As a rule, the solving of every problem is not an isolated process but is united with the previous experience. In such a case the integrity of the solving process perception appears.

The first recommendation may be the following.

- Wait while the solution would come into your mind itself.

Only if you wait too long then you can go to the next recommendation.

The next step of decision making is a switch-over to an over-goal.

- Why I try to solve exactly this problem? May be it is better to read some other book or to go for a walk.

Then the following two alternatives are possible.

- To leave alone this problem and to find a new one.
- Has the problem in such a setting some solution? Is it probable that I would find a solution?

If nevertheless you are planning to solve the problem then begin the decision with the end.

- Present obviously what must you do.

Use the next recommendation until you feel that it can bring some new idea of a decision.

- Use the language of draughts, formulas, algorithms or programs which allows to reformulate the problem. What does become more clear after the reformulation?
- Simplify the problem.

Increase the volume of input.

Add an additional premise to the conditions.

Decrease the number of unknowns.

Weaken the conclusion.

Specialize the problem.

- Formulate an intermediate goal or problem. It is desirable that the intermediate one be sufficiently far from the data and unknowns of the main problem.

Actually such an intermediate problem is a “mathematical brick” named so by V.A. Ufnarovskiy [5].

- Find an almost solved similar problem, a solved problem with similar output or conclusion, a solved problem with similar input.
- Sort the information about the problem according to its importance and usefulness.
- Order the discovered difficulties according their significance.
- Introduce auxiliary elements or new dimension, pass to an over-problem. Generalize!

Sometimes persistence in going to the goal does not permit to see slight circumstances which are key ones in reaching the goal. Moreover, sometimes a key role of some circumstance is already in the subconsciousness. Hence, there appear the next recommendation.

- Fix the appeared thoughts. If you fill something like “This detail promises something”, then it is useful to write down a short denotation of this detail and its connection with the proper of investigation.

As a rule it is not needed to make big efforts in this direction just now. Further sharpening of the direction may appear itself.

- Formulate the problem in metafora language! Search for analogy! It may lead to discovering new facts.
- Make a list of the simplest particular and limiting cases. Order them according to their clarity and studying!
- Look for discovering of some often repeated fact or scheme.
- Control your observation by means of a luck thought or idea.
- Check your supposing by means of particular cases and facts that follow from it.
- Use the symmetry of the problem. For example, use the principle of sufficient basis proposed in the book [4].
- Make the problem more precise! Or make it less precise! (Change the degree of precision.) Change terms by their definitions!
- Reduce the problem to the one from another mathematical theory. Formulate it in the language of logic, algorithms, programs. In the language of variables and functions. Pick out a parameter and a proposition for mathematical induction implementation.
- Reduce the problem to itself. Use the recursion or the induction.
- Reduce the problem to the conjunctive normal form only with universal quantifiers before it. Remove surplus elementary disjunctions. Find locus for every elementary disjunction and then find an intersection of all these loci. In the other words decide the problem as a search one.
- Reduce the problem to the disjunctive normal form only with existence quantifiers before it. Remove surplus elementary conjunctions. In the other words, try to decide the problem as a proof one. Locate the main case (main elementary conjunction).
- Beginning with the main case try to receive the general solution with the help of superposition of special cases.

The decision of a search problem is provided by more formalized objects than a decision of an existence one.

- Introduce a common notation for a sub-problem which appears at least twice (may be with different parameters).
- Apply trial and error method of G. Polya [4] for searching among small number of formalized objects with easy enough checking if it is a solution.
- Check a consistency of some conditions. Of all conditions!
- Check an independence of every condition from the other problem conditions.

- Sort out and use your knowledge.
- Sort out and use your abilities.
- Adapt abilities to the deciding problem.
- If you have no idea of a decision then return to definitions (point 20).

Of course, there is a lot of problems which does not satisfy any of these recommendation. If you solve such a problem try to formulate a new recommendation.

- Include the idea of the solved problem into your experience (base of tools, base of knowledge, base of solution search).

What is the main sense of the problem?

What was the most important idea in the process of decision?

What was the main difficulty?

What could I do better?

This detail I have mist. What must be the peculiarity of an intellect permitting to see this detail?

Whether there is a method which I can use the next time in a similar situation?

What else approaches to this problem decision are there?

Bibliography

Altshuller G.S. Creativity as an exact science. M.: Soviet radio. 1979. 175 p. (In Russian)

Kosovsky N.K. Technology of innovation development. St.Petersburg: St.Petersburg state university. 2013. 186 p. (In Russian)

Polya G. How to solve it. Princeton University Press. ISBN 0-691-08097-6.

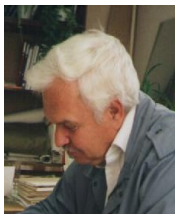
Polya G. Mathematical discovery.

Ufnarovsky V.A. Mathematical aquarium. Kishinev: Shtiintsa. 1987. 216 p. (In Russian)

Acknowledgement

The paper is published with financial support of the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Authors' Information



Nikolay Kosovskiy – Dr., Professor, Head of Computer Science Chair of St.Petersburg State University, University av., 28, Stary Petergof, St.Petersburg, 198504, Russia,

e-mail: kosov@NK1022.spb.edu

Major Fields of Scientific Research: Mathematical Logic, Theory of Complexity of Algorithms.

ANALYSIS OF FEATURES AND POSSIBILITIES OF BANK FUNCTIONING EFFICIENCY BASED ON THE METHOD OF STOCHASTIC FRONTIERS

Oleksandr Kuzomin, Vyacheslav Lyashenko

Abstract: *Given paper shows the importance of analyzing the banks performance in the process of their operation and development. The place and the role of researching of stochastic frontiers method for the banks performance assessment have been defined. The features and the possibility of assessing efficiency of banks activity with the usage of the of the stochastic frontiers method have been studied. The examples of performance assessment of banks activity in the aspect of the credit extension have been given. The expediency of the consideration of the multiple stochastic frontiers in the assessment of the effectiveness of the banks activities has been proved.*

Keywords: *efficiency, stochastic frontier analysis, financial flows, loss function, technical efficiency, allocative efficiency.*

ACM Classification Keywords: *G.3 Probability and statistics – Correlation and regression analysis, Multivariate statistics; J.1 Administrative data processing – Financial, J.4 Social and behavioral sciences – Economics.*

Introduction

Regularity of financial flows movement in the system of market interaction of inter-action between the different economic entities (including such specific as the state) to a large extent supported by the banking sector of the economy. This is due to the fact that the banking sector of economy contributes to the uncommitted resources transfer, not only from one economic entity to another, but also from the people to the various sectors of the economy by means of the savings transformation in investment re-source. Therefore, the investigation of the efficiency of banks activity on the basis of the analysis of their financial flows movement is under direct interest of researchers. The importance of the carrying out a relevant analysis is attributed not only to the possibility of studying of the bank financial flows impact on the development of economic stability overall, but also on the Bank and the banking sector development [Kuzemin, Lyashenko, 2009]. Thus, the performance analysis of banks can be considered as one of the priorities in researches concerning the economic dynamics and economic growth.

Performance analysis of banks and method of stochastic frontiers:

When considering the effectiveness of the bank activities primarily are focused on the analysis of various indicators of the operation results and the banks development in terms of the various financial flows movement.

For example, in the work [Collier, McGowan, Muhhamad, 2006] and in the work [Aarma, Vainu, Vensel, 2004] the performance analysis of banks is made based on decomposition of return on assets and return on equity of banks under study.

However, in the work [Williams, 2005] the performance analysis of banks is based on econometric methods. The essence of this analysis consists of the consideration of the mutual influence of the various financial flows movement on the assessment of the Bank activities. Given assessment reflects the various performance and development indicators as unique bank, and whole sector.

Stochastic frontier analysis (SFA) is one of the most prospective analyses based on econometric methods among the approaches of the assessment of efficiency of banks activity. The essence of the SFA is described in works [Farrell, 1957], [Aigner, Lovell, Schmidt, 1997], [Battese, Coelli, 1992] and based on:

An efficiency frontier construction with the methods of statistical analysis,

Positioning of studied economic process or object relative to the resulting efficiency frontier,

Effectiveness determining of the studied economic process or object to a function that describe the attainability of constructed efficiency frontier.

The following model is used for of the efficiency frontier formalization [Aigner, Lovell, Schmidt, 1997]:

$$y = f\left(x, \beta\right) + \varepsilon \,, \tag{1}$$

$$g = v - u \,, \tag{2}$$

where y – vector of outcomes of the of the studied object or process (in this case the results of banks activities),

x – a vector of the resources used to produce certain banks results;

f – a function of the banks efficiency frontier;

β – a vector of function parameters of;

ε – a composite random element of the model;

v – a vector of random vibration of a model;

u – a vector describing the technical inefficiency of banks.

Then the effectiveness of a particular bank $\left(i\,, i = \overline{1,l}\,, l\right.$ – a total number of studied banks), or rather its technical efficiency (TE_i), can be calculated as follows [Johdrow, Lovell, Materov, Schmidt, 1982]:

$$TE_i = e^{-M\left(u_i \mid \overline{\varepsilon_i}\right)} \,, \tag{3}$$

where $M(u_i \mid \overline{\varepsilon_i})$ – a conditional mathematical expectation of at an estimated value of $\overline{\varepsilon_i}$. In total $M(u \mid \varepsilon) = r(\varepsilon)$ is a regression vector of u by ε.

Thus the construction of a function that describes the efficiency frontier of banks and the calculation of this efficiency greatly depends on the following items:

Select of the model function form that describes the frontier of the bank effectiveness, there are various options for preferences that take into account both economic substance and significance of the models,

Select of u – the distribution component of the bank technical inefficiency, ε – a compound random element of the model, where, as a rule is considered a half-normal, truncated and exponential distribution [Murillo-Zamorano, 2004]. Although in general, the distribution component can be of any kind, but must be taken into ac-count non-negative.

It should be added that the loss function ε selection in considering the model function form that describes the efficiency frontier of the bank activity has a definite impact on the estimated value , and therefore on the value of the individual components of the vector u.

Assessment of the bank's performance with different loss function selection in the description of the functions model form that describes its frontiers:

Let make a performance analysis of Ukrainian banks in the area of the extension of credit recourses, as a specific example that shows the effect of the loss function selection in considering model function forms that describes the frontier of the bank activity. This is due to the fact that the credit activity is one of the main components of the

banks functioning. The selection of the Ukrainian banks as objects for case study is associated with the following facts:

On the one hand, the performance rating of Ukrainian banks on the basis of analysis of stochastic frontiers is not considered in related studies,

On the other hand, crediting problems of real sector of economy are one of the vital issues of development of the banking sector of the Ukrainian economy at the present stage of progress.

For further analysis let consider the production Cobb-Douglas function as a model function form that describes the frontiers of performance of banks in terms of extension of credit recourses. The basic parameters of this function are:

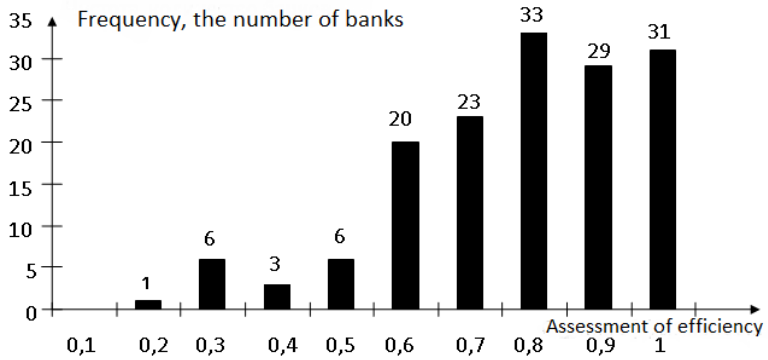
the value of total volume of credits to various business entities and the people (KR) is taken as a vector of evaluation banks performance, the value of the volume of resources attracted from other banks (DB), the value of the volume of deposits attracted from economic entities and people (DO), the value of the volume of administrative and other operating expenses of banks (AZ) are taken as separate components of the vector of resources used to produce certain results of the bank. (DO) represents the ability of banks to use financial resources to extend credits, (AZ) displays the fact of labor resources usage for the banking activities

Then the form of the corresponding function model that describes the frontiers of the effectiveness of the banks activities takes the following form in the aspect of the extension of credit resources:

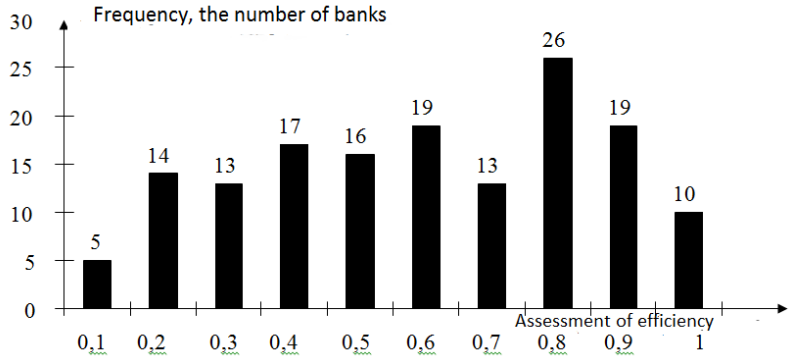
$$\mathrm{Ln(KR)} = \beta_0 + \beta_1 \cdot \mathrm{Ln(DB)} + \beta_2 \cdot \mathrm{Ln(DO)} + \beta_3 \cdot \mathrm{Ln(AZ)} + \varepsilon. \quad (4)$$

Appropriate values of the Ukrainian banks activities as of the date 01.04.2010 (the data for analysis were taken from the site www.bank.gov.ua) have been examined with the purpose of researching the influence of the loss function selection in considering forms of the model function that describes the frontiers of the bank effectiveness. Two methods of the loss function calculation have been studied in given paper, they are ordinary least squares and maximum likelihood method. In both cases, statistically significant models describing the frontiers of the effectiveness of the banks activities in the aspect of the credit resources extension requested in accordance with the Cobb-Douglas function have been obtained. Nevertheless, the results evaluating the performance of the studied banks in terms of the credit extension are not identical.

The histogram of the assessment of the effectiveness of the credit assignment for the form of model function is presented on the following picture. This histogram describes the limits of the efficiency of banks' activity, where the function of loss is calculated by the ordinary least squares method.



Next picture shows a histogram of evaluating the effectiveness of credit extension for the form of the model function that describes the frontiers of the effectiveness of the banks activities, where the loss function is calculated by the maximum likelihood method.



As can be seen from the data presented on both histograms, evaluations of the performance of the banks under consideration in terms of the credit extension are different. This is determined by the loss function used in determining the parameters of the model function that describes the scope of performance of banks.

Thus objective question about the features of the method of analysis of the effectiveness of the banks, based on method of investigation of stochastic frontiers there is raised. The answer to this question lies in the fact that method for evaluation of the performance of banks for the application of stochastic frontier analysis may be useful, first of all, to distinguish nature of determination of the appropriate ratings.

So when evaluating the performance of banks will be cover only a specific time slot and a specific form of the model function that describes the frontiers of the effectiveness of the banks, the selection of the loss function should be based on the most important statistical estimates in the model function that describes the boundaries of performance of banks.

If the evaluation of the effectiveness of the bank will be conducted in a comparative aspect, it is natural to use the same models, which determine the form of the function describing the frontiers of performance of banks for different data. In such case, it is important to show the obtained statistical evaluation forms of the model function that de-scribes the frontiers of the effectiveness of the banks when selecting a particular loss function. In other words, the manifestation of the features of the method of analysis of stochastic frontiers to calculate estimates of the effectiveness of the banks is primarily observed in a comparative analysis of the functioning of banks for different data sets. But exactly a comparative analysis is the basis for making the right decisions on the functioning and development of banks. Thus, for a comparative analysis for evaluating the performance of banks is important to show the most significant forms of estimates of the model function that describes the effectiveness of the frontiers of banks activities. At the same time, in order to obtain more reliable estimates of such a frontier a collection of primary data used for analysis can be modified. Objective factors of such variation may be the selection of particular set of banks used for further study.

In this case, the important fact is that some of the banks in different intervals can be either under the interim administration, or deprived of licenses for certain activities, which imposes restrictions on the formation of a sample of banks used for further studies. However, taking into account these factors can significantly affect the degree of reliability and relevance of estimations in selecting the form of the model function that describes the frontier of performance of banks.

The use of the stochastic frontiers set as an opportunity to expand the analysis of the effectiveness of the banks:

It is also important to consider opportunities offered by the analysis of the effectiveness of the banks based on the use of the stochastic frontiers set, along with its certain features.

To our opinion, one of the key features of the use of the stochastic frontiers research, for analyses of the banks' work, is to consider these assessments for a variety of functions describing the efficiency frontiers of banks work with the accounting of the economic substance diversity.

In particular, from the point of view of providing of credit resources, such diversity involves not only the consideration of the banks' work effectiveness, but also the consideration of the amount of credits given to business entities and people.

Then, along with equation (4) it is also appropriate to consider the following equations, which define the models describing effectiveness frontiers of banks' work:

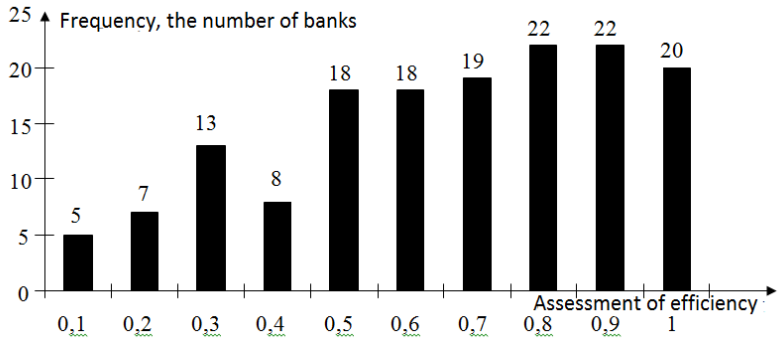
$$Ln(KRS) = \beta_{01} + \beta_{11} \cdot Ln(DB) + \beta_{21} \cdot Ln(DO) + \beta_{31} \cdot Ln(AZ) + \varepsilon_1 \quad (5)$$

$$Ln(KRN) = \beta_{02} + \beta_{12} \cdot Ln(DB) + \beta_{22} \cdot Ln(DO) + \beta_{32} \cdot Ln(AZ) + \varepsilon_2 \quad (6)$$

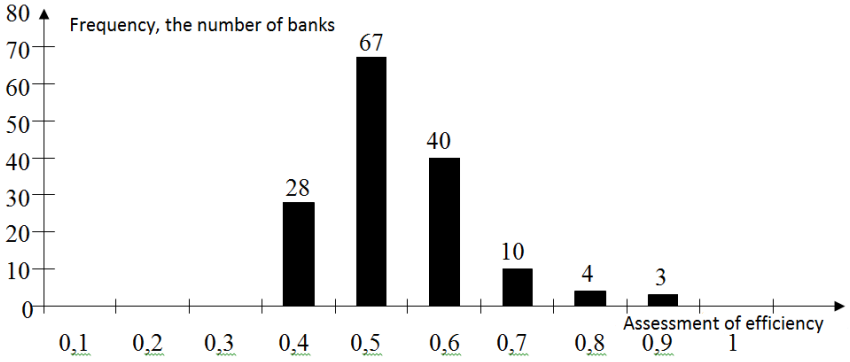
where KRS – the amount of credits given to various business entities, KRN – the amount of credits given to people.

Own values of the effectiveness assessments are calculated, according to the models, that reflect certain assignment of frontiers of banks' work effectiveness.

Next picture shows a histogram of the assessment of the effectiveness of providing a credit to business entities for the given form of the model function. This form describes the frontiers of the efficiency of banks' work with the function of the loss calculated by the method of maximum likelihood.



Next picture shows a histogram of the assessment of the effectiveness of providing a credit to people, for the given form of the model function. This form describes the frontiers of the efficiency of banks' work with the function of the loss calculated by the method of maximum likelihood.



As can be seen from three previous pictures assessments of the effectiveness of the banks' work in terms of credit providing are different. Moreover, three previous pictures show that the effectiveness of the banks' work in terms of credit providing (for a certain period of time) is largely determined by the efficiency of credit resources

given different business entities and to they are in a positive correlation. In this case, for considered banks, the correlation between the efficiency of providing a credit in general and the effectiveness of providing a credit to people is negative. The above allows making a conclusion about the possibility of establishing the relationship between the models of functions, describing the effectiveness of the relevant frontiers of banks in terms of providing a credit. In addition, for comparison of such correlation it is important to consider the transformation between the technical and allocative efficiency, where the last reflects the efficiency of available resources [Cooper, Seiford, Tone, 2007], that is especially important in terms of providing a credit.

It is based on the fact that the use of resources for providing a credit may not be effective in the case of insufficient use of all available resources for crediting, and vice versa not appropriate resources can be used for lending operations (such as time limits for applying the deposits), that leads to unreasonable providing of non-existent resources.

In general, the effectiveness of the work of l researched banks in terms of credit providing in whole (TE^o) and the effectiveness of the work of l banks providing a credit to different business entities (TE^s) and to people (TE^n) can be presented in its simplest form, in the form of a regression relationship:

$$TE^o = \lambda_0 + \lambda_1 \cdot TE^s + \lambda_2 \cdot TE^n + \eta \tag{7}$$

where λ – parameters of the regression model;

η – random member of the model.

If we take into account the formulas (3-6), model (7) can also be written as follows:

$$e^{-M(u|\bar{\varepsilon})} = \lambda_0 + \lambda_1 \cdot e^{-M(u_1|\bar{\varepsilon}_1)} + \lambda_2 \cdot e^{-M(u_2|\bar{\varepsilon}_2)} + \eta \tag{8}$$

At the same time, the ratio between the individual assessments of the bank's work can be a reflection of the ratio of the results of the bank activity, that exist during consideration of such assessments. In particular, in this case above, the correlation between the estimates of the effectiveness of the bank in terms of lending to various business entities and the public in some way reflects the existing ratio in the volume of loans as a business entity, and the public. In particular, in the case above, the correlation between the assessments of the effectiveness of the banks' work, in terms of providing a credit to various business entities and to people, in some way reflects the existing ratio in the amount of credits given to business entities, and people. This statement follows from equation (3), considering that value ε can be represented as a function which depends on the amount of credits in accordance with equations (4) and (5).

Conclusion

Thus, some of the features and possibilities of the analysis of the effectiveness of the banks' work based on stochastic frontiers research were considered in this paper. These features and possibilities provide more objective results. The selected features of the application of the method of investigation of stochastic frontiers of the analyses of the efficiency of banks are considered by other authors. But it should be emphasized that, we have focused on the account of such features in terms of the comparative analysis of the evaluation of banks' work. Allocation of such direction in the analysis of the effectiveness of the banks' work is chosen due to the fact, that features of stochastic frontiers method appear precisely within the comparison. At the same time, mutual consideration of several stochastic frontiers for the comparative analysis of the banks can be considered as a new perspective of application and development of stochastic frontiers method in the practice of economic analysis. It is important to find the correlation between the technical and allocative efficiency. It allows us to

consider the overall economic efficiency of the researched events, processes and objects at a different level of understanding of individual kinds of efficiency. he questions of formalizing relations between the assessments of the efficiency of different banks from the point of their various activity directions aren't less important. This formalizing allows not only to extend the appropriate analysis, but also refines it. It helps to uncover and to take into consideration the correlation of stochastic frontiers.

Acknowledgement

The paper is published with financial support by the project ITHEA XXI of the Institute of Information Theories and Applications FOI ITHEA (www.ithea.org) and the Association of Developers and Users of Intelligent Systems ADUIS Ukraine (www.aduis.com.ua).

Bibliography

[Aarma, Vainu, Vensel, 2004] Aarma A., Vainu J., Vensel V. Bank Performing Analysis: Methodology and Empirical Evidence (Estonian Banking System, 1994-2002). – February 6, 2004 // http://ssrn.com/abstract=499434.

[Aigner, Lovell, Schmidt, 1997] Aigner D. J., Lovell C. A., Schmidt P. Formulation and Estimation of Frontier Production Function Models // Journal of Econometrics. – 1997. – № 6.

[Battese, Coelli, 1992] Battese G. E., Coelli T. J. Frontier Production Functions, Technical Efficiency and Panel Data: With Appli-cation to Paddy Farmers in India //Journal of Productivity Analysis. –1992. – № 3.

[Collier, McGowan, Muhhamad, 2006] Collier H.W., McGowan C.B., Muhhamad J. Financial analysis of finan-cial institutions in an evolving environment / Proceedings of the Meeting of the Decision Sciences Institute. – Oklahoma City, March, 2006.

[Cooper, Seiford, Tone, 2007] Cooper W. W., Seiford L. M., Tone K. Data Envelop-ment Analysis. A comprehensive text with models, applications, references and DEA-solver software. – Springer, 2007.

[Farrell, 1957] Farrell M. J. The Measurement of Productive Efficiency // Journal of the Royal Statistical Society. – I957. – ACXX. – Pt. 3.

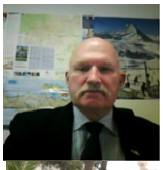
[Johdrow, Lovell, Materov, Schmidt, 1982] Johdrow J., Lovell C. A., Materov I. S., Schmidt P. On the Estimation of Technical Inefficiency in the Stocha-stic Frontier Production Function Model //Journal of Econometrics. – 1982. – № 19.

[Kuzomin, Lyashenko, 2009] Kuzomin A., Lyashenko V. Methods of comparative analysis of banks functioning: classic and new approaches // International Journal In-formation Theories & Applications. – 2009. – Volume 16. – № 4.

[Murillo-Zamorano, 2004] Murillo-Zamorano L. Economic Efficiency and Frontier Techniques // Journal of Economic Surveys. – 2004. – Vol. 18.

[Williams, 2005] Williams J. Financial Liberalisation, Crisis, and Restructuring: A Comparative Study of Bank Performance and Bank Governance in South East Asia // Journal of Banking & Finance. – 2005. – № 29.

Authors' Information



Oleksandr Kuzomin *- Doctor of Technical Science; Information Science 14, Lenin Ave., 61166, Kharkiv, UKRAINE**;** Tel/fax: +38(057)7021515 ;*
mailto:kuzy@daad-alumni.de

Vladislav Lyashenko *- Senior Researcher 14, Lenin Ave., 61166, Kharkiv, UKRAINE; Tel/fax: +38(057)7021515;*
mailto:kuzy@kture.kharkov.ua

PECULIARITIES OF LINKED DATA PROCESSING IN SEMANTIC APPLICATIONS

Sergey Shcherbak, Ilona Galushka, Sergey Soloshich, Valeriy Zavgorodniy

***Abstract**: Nowadays linked data popularity increases along with information description in the form of RDF-triplets. Efficient implementation of users' interaction with these kinds of data requires studying communication procedures with triple stores. One of the main difficulties, that are currently unsolved, is the complexity of dynamic querying procedures. We try to deal with this issue by creating query patterns and decreasing query complexity. A unified search interface is developed, which enables visual querying the triple stores implemented through OpenLink Virtuoso universal server. Visual queries are automatically converted into SPARQL query language, which is used for accessing the triple stores. After the query is executed, a user gets the desired context with triplets according to constraints for predicates and objects. Also the formal model is developed, which mathematically describes linked data based on partially defined object schemas. Existing search model for distributed environments is improved. The developed practical implementation for medical institution is under stage of manufacturing application.*

***Keywords**: linked data, RDF data management, query pattern, triple store, semantic application.*

***ACM Classification Keywords**: E.2 DATA STORAGE REPRESENTATIONS (Linked representations), I.2.4 Knowledge Representation Formalisms and Methods (Representation languages)*

Introduction

The development of linked data standards (like "RDF Primer", "RDF/XML Syntax Specification" "SPARQL Protocol for RDF", "SPARQL Query Language for RDF" and other W3C standards) and their support by authoritative software development companies determine trends of evolution for the global network as a huge data and knowledge storage, which provides means for accessing structured data through specialized search interfaces [Ma, 2009]. Linked data are defined using Resource Description Framework (RDF) [Bizer, 2007]. This definition takes place in a form of triplets (subject – predicate – object) or quads (named graph – subject – predicate – object). For simplicity we shall use the term "triplet" to define triplets as well as quads, if it does not lead to contradictions.

RDF model assumes distributed storage of objects along with their schemas (if such schemas exist) on different web-servers. Such servers include integrated or external triple stores. SPARQL (recursive acronym for SPARQL Protocol and RDF Query Language) is used for accessing triple stores. If we draw analogy with relational database, SPARQL is alike to SQL in a way [DuCharme, 2011]. To use such a language of querying effectively, one should possess some special knowledge and skills. It neither encourages SPARQL popularization, nor increases triple store expansion.

Our goal is to increase the efficiency of search based on triple stores by means of decrease in complexity of SPARQL dynamic querying and due to implementation of query patterns for further search interface development. On our way to it, we analyze current approaches and possibilities of linked data in the context of search, modern user interface development techniques for linked data storages. In the next sections,

formalization of linked data and search procedures are provided, architecture and unified search interface based on SPARQL are described.

Search interfaces for triple stores

Linked data search is characterized by determination of objects that belong to some application domain. In case of using such an approach, the content of documents is presented as a collection of objects grouped together by certain context [Allemang, 2009]. To determine object context we shall use a term “graph” that belongs to a quad. Information search assumes identification of named graph triplets and context identification that match user specific criterions or restrictions.

Implementation patterns should be designed for efficient search implementation and friendly user interfaces [Beck, 2008]. The patterns should provide visual querying to named graphs of triple stores without requiring any knowledge or skills in SPARQL from users. Fig. 1 shows architecture of a typical application on triple stores and SPARQL [Kalfoglou, 2009].

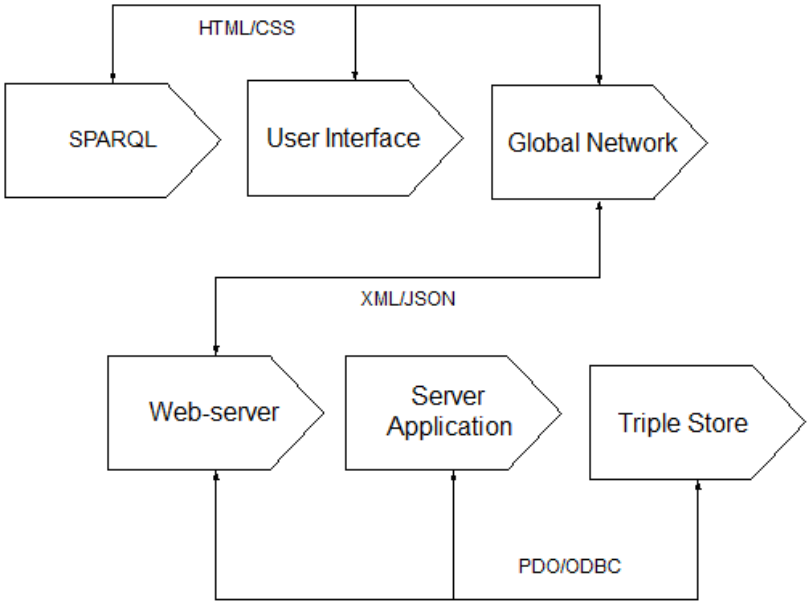


Figure 1. Architecture of a typical application with a triple store

User’s query on SPARQL is transmitted from web-application (with front end on HTML and CSS) to the triple store, where queries are processed using interfaces PDO and ODBC. Processing results are sent back to users in XML or JSON format. SPARQL queries are traditionally written by users. This requires some specialized knowledge about SPARQL and structure of objects residing in triple store. One of applications, that realize such a functionality, is ISQL web-interface from OpenLink used in Dbpedia. This application possesses universal data accessing techniques, and it is characterized by operation stability.

Let take a look at the proposed solution which defines user interaction with triple stores through SPARQL (see fig. 2). A user gains access to named graphs of triple store through visual SPARQL query builder. It gives an opportunity to achieve information about graph topology automatically. This information includes predicates and object data types. It can be used to create search request through determination of user restrictions (filters) on data returned from a triple store.

Denote user restrictions as a set of rules with conditions for objects returned from a triple store. Object type defines semantics of implemented comparison operation, i.e. the way query builder reacts on operation signs.

The following basic operations can be used to return objects: "=" (equality), ">" (more), and "<" (less). Equality sign for integer (xsd: integer) objects and float (xsd: float) objects, as well as equality sign for string objects (xsd: string), means that the analyzed objects are equal. Signs ">" and "<" for integer and float objects have their direct meaning also, whereas ">" and "<" signs for string objects mean that one object is a substring of another.

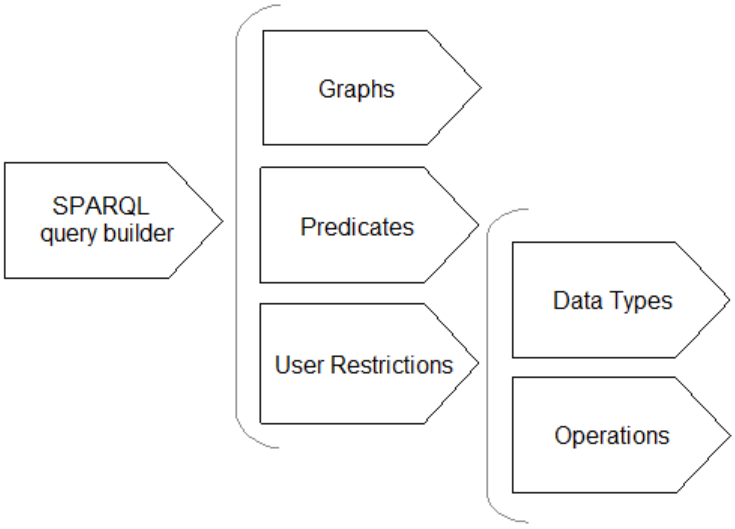


Figure 2. Schema of visual querying to triple stores

Consequently, a user may create rules to return content from a named graph of triple store. Typical SPARQL query to triple stores can be created automatically according to obtained rules. Practical aspects of such querying are observed in a section, and in the next section a formal presentation of triplets and linked data is given.

Formal model of linked data on partially defined schemas

Linked data can be presented in a following way:

$$t = < g, s, p, o >, \tag{1}$$

where t is a triplet, g – named graph, s – subject, p – predicate, and o – object.

A collection of structures t, defined by formula (1), we denote as a triple store:

$$T = \{t_i\}, i = \overline{(1, n)}, \tag{2}$$

where t_i is the i-th triplet and n – the number of triplets in a store.

Context information is needed to be assumed for search operation. The context is considered to be identical for all t_i with the same g. The following formula can be used to define context:

$$G = \{g_j\}, j = \overline{(1, m)}, \tag{3}$$

where G is a set of all contexts from the store, g_j – j-th context of triple store, and m – number of contexts from the store.

Associate each context with the collection of tree elements ($< s, p, o >$) from formula (1). Consequently, the context can be defined as follows:

$$\forall g \in G:\ g = < S, P, O >, \quad (4)$$

where S is a set of subjects, P – a set of predicates, O – a set of objects.

An object can be defined as follows:

$$\forall o \in O:\ o = < T, L, V >, \quad (5)$$

where T – data type, L –language of presentation, V – value.

Assuming linked data peculiarities (namely optionality to define schemas, data types, languages used for object value presentation), we consider object elements optional, and object schema (according to formula (5)) is considered as partially defined by formulas (1)–(4).

To implement search tasks on G, we shall modify formula (4) by adding new element F for a function set. These functions may be performed on sets of context elements G.

$$\forall g \in G:\ g = < S, P, O, F >, \quad (6)$$

Practical aspects of function set implementation on SPARQL are described below. If it does not lead to contradictions, terms “context” and “named graph” are further used as synonyms.

Practical aspects of search interface implementation

Typical application with a triple store (showed in fig. 1) and visual querying (showed in fig. 2) assume several actions to be performed on SPARQL for dynamic return of named graph topology. Such actions include: returning a list of named graphs from a triple store, returning a list of named graph predicates from a triple store, returning predicate data type. These SPARQL queries may be more complicated, for instance queries that return some named graph or group, which name equals (or partially corresponds) to a predefined criterion. Language tags can also be used as user restrictions. In this case, a filter can be set to return objects from a graph in Russian. Several restrictions can be used simultaneously to search for an object [DuCharme, 2011].

The proposed example of user interface is interesting as ontology integration on the user interface layer is quite a novel field of research [Paulheim, 2011]. Fig. 3 shows search interface for the queries mentioned above. It is implemented on PHP and tested on OpenLink Virtuoso server. OpenLink Virtuoso is chosen as a triple store, reasoner, RDF generator and SPARQL endpoint. It supports direct mapping and many programming languages, including C, C++, Python, PHP, Java, Javascript, C#, ActionScript, Tcl, Perl, Ruby, Obj-C [Segaran, 2009, Hitzler, 2009].

The proposed search interface provides users with additional information that can ease query building and decrease time required for it. Query building procedure reduces to selection of named graphs (which are interesting for users) and setting restrictions on predicates. Search results are grouped according to the subject in a tabular style. Consequently, data are presented at a clients’ front end in a friendly way, and users may know nothing about SPARQL query existence.

Although RDF language gives an opportunity to create graph structures of arbitrary level of complexity [Powers, 2003], there are several restrictions for communication with a triple store through ODBC (Open Database Connectivity). Data in a triple store are structured according to object-oriented principles, i.e. named graph is a container (class) for a set of objects belonging to it. The objects are uniquely identified by a triplet subject. This restriction provides more possibilities for software developers that use relational databases. If we do not take

these restrictions into account, the proposed method will work, but data generation sense will be changed. To consider object restrictions, named graphs based on typical SPARQL queries are proposed to be created.

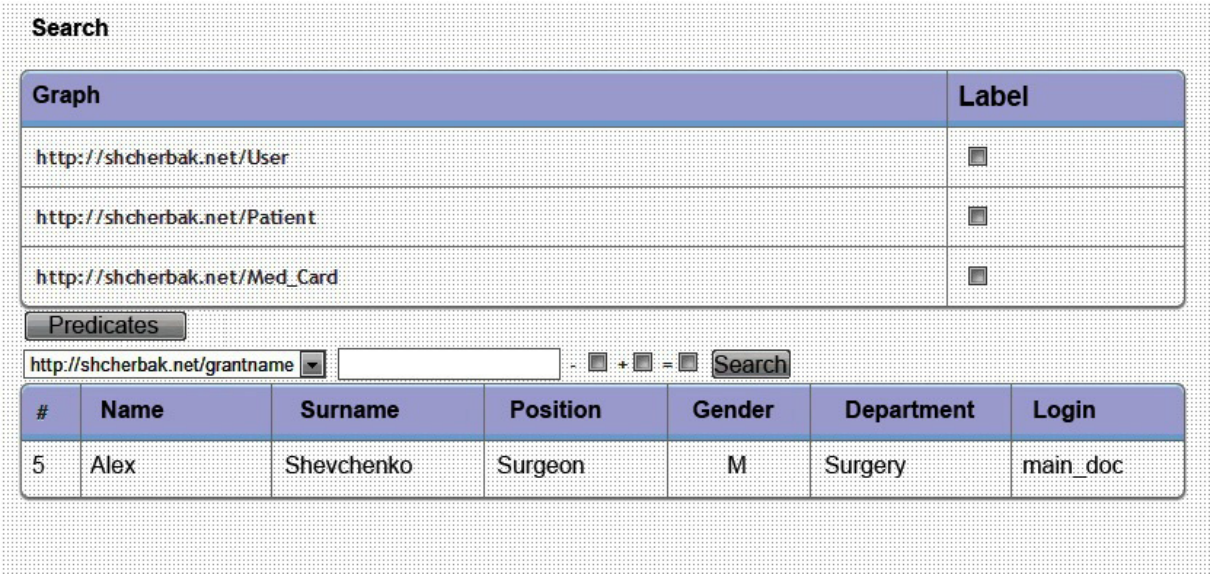


Figure 3. User interface with search results in a tabular style

Conclusion

The proposed architecture and example of visual SPARQL querying implementation is oriented on fast input of queries to data stores and search quality perfection. It permits searching in conditions of partially defined object schemas. Linked data model is offered for partially defined schemas. Search model for distributed environments with partially defined schemas is extended. Technological recommendations are suggested for implementation of user interfaces to triple stores with their automatic generation. After a query is executed, a user obtains context with triplets that match to restrictions on predicates and objects set by this user.

Practical implementation is performed on PHP and tested on multi-model data server OpenLink Virtuoso. This server is selected because of several reasons. First of all, it is cross-platform and can be used for relational data management as well as RDF and XML data management, free text content management and full text indexing. Secondly, it supports lots of programming languages and semantic web technologies, and thirdly, it is available for free and commercial use. The proposed models and technologies are highly efficient in a sense of enterprise market appeal from the point of basic principles suggested in [Wood, 2010]. They are also interesting from the point of modern effective solutions for semantic applications.

Acknowledgements

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

[Allemang, 2009] D. Allemang, J. Hendler. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. Morgan Kaufmann, 2009, 352 p.

[Beck, 2008] K. Beck. Implementation Patterns. Addison-Wesley, 2008, 157 p.

[Bizer, 2007] C. Bizer, R. Cyganiak, T. Gauß. The RDF Book Mashup: From Web APIs to a Web of Data. In: 3rd Workshop on Scripting for the Semantic Web, Vol. 248, 2007, 6 p.

[DuCharme, 2011] B. DuCharme. Learning SPARQL. O'ReillyMedia, 2011, 258 p.

[Hitzler, 2009] P. Hitzler, M. Krötzsch, S. Rudolph. Foundations of Semantic Web Technologies. Chapman and Hall/CRC, 2009, 456 p.

[Kalfoglou, 2009] Y. Kalfoglou. Cases on Semantic Interoperability for Information Systems Integration: Practices and Applications. IGI Global, 2009, 350 p.

[Ma, 2009] Z. Ma, H. Wang. The Semantic Web for Knowledge and Data Management: Technologies and Practices. IGI Global, 2009, 367 p.

[Paulheim, 2011] H. Paulheim. Ontology-Based Application Integration. Springer, 2011, 270 p.

[Powers, 2003] S. Powers. Practical RDF. O'ReillyMedia, 2003, 352 p.

[Segaran, 2009] T. Segaran, C. Evans, J. Taylor. Programming the Semantic Web. O'Reilly Media, 2009, 302 p.

[Wood, 2010] D. Wood. Linking Enterprise Data. Springer, 2010, 291 p.

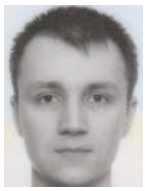
Authors' Information

Ilona Galushka *– postgraduate student of Information and Control Systems department in Kremenchuk Mykhailo Ostrohradskyi National University, P.O. Box: 39600, Ukraine, Kremenchuk, Pershotravneva Street, 20; e-mail: anoli@gmail.com*

Major Fields of Scientific Research: Linked data, agent technologies

Sergey Shcherbak *– PhD of Information and Control Systems department in Kremenchuk Mykhailo Ostrohradskyi National University, P.O. Box: 39600, Ukraine, Kremenchuk, Pershotravneva Street, 20; e-mail: ontolog@gmail.com*

Major Fields of Scientific Research: Semantic web, web services, RDF data management



Sergey Soloshich *– postgraduate student of Information and Control Systems department in Kremenchuk Mykhailo Ostrohradskyi National University, P.O. Box: 39600, Ukraine, Kremenchuk, Pershotravneva Street, 20; e-mail: soloshich@gmail.com*

Major Fields of Scientific Research: Linked data, agent technologies

Valeriy Zavgorodniy *– senior lecturer of System Software department in Dneprodzerzhinsk state technical university, P.O. Box: 51900, Ukraine, Dneprodzerzhinsk, Dneprostroevska Street, 2; e-mail: valera_ddtu@i.ua*

Major Fields of Scientific Research: Linked data, agent technologies

KEY FRAME PARTITION MATCHING FOR VIDEO SUMMARIZATION

Olena Mikhnova, Nataliia Vlasenko

Abstract: *Summarization of video content is a complex task that requires feature selection and frame matching. To extract meaningful frames, named key frames, we have proposed partitioning of frames with Voronoi diagrams for further region matching throughout the video sequence. A unique partition metric has been used that takes into account color and textural, structural and geometric properties of Voronoi regions. Feature set designed for CBIR and CBVR has been analysed. The reasonable feature selection and incorporation into frame matching process has permitted to obtain competitive results. All these actions have allowed revealing significant changes in content while omitting slight deflections and repeats. Key frame extraction procedure has been described in detail. The proposed method has been checked on different test samples and compared with existing methods for precision and recall.*

Keywords: *key frame extraction, generator point detection, Voronoi diagram, partition metric, feature set.*

ACM Classification Keywords: *I.2.10 Vision and Scene Understanding (Video analysis), I.4.6 Segmentation (Edge and feature detection)*

Introduction

Multimedia utilization has greatly increased during the last decade, the amount of digital libraries had grown to enormous sizes and users started requiring instruments to deal with these data. Video is the most informative type of multimedia, as it consists of audio and graphical information simultaneously. Moreover, this information dynamically changes in time, which is one of the main difficulties for processing and analysis. Video summarization is among video recognition tasks that still lacks in performance and accuracy of computational procedures [Sonka, 2007].

Video summarization techniques are engaged in archiving, browsing and searching, cataloging and indexing, as well as improvement of information overload. It aids in maintenance of usability and accessibility to stored videos. Summarization may be of two types: static and dynamic. The last one is usually called skimming. The product of video summarization is a set of meaningful static images that depict video content, while the product of skimming is a shortened video [Laganière, 2008]. The subject of our research is the first type of summarization, static key frames.

The origin of video summarization and skimming comes from movie editing, when a film director decides which frames should be cut off. The relation of initial material to the resultant is usually not bigger than 20:1 [Rubin, 2005]. There are many movie editing techniques, but not a single rule of doing so. Acceptance of some frames and omission of others is a point of individual, cultural, or even political taste of an editor [Goldman, 2007]. American film director and scenario writer, John Huston stated that video editing should omit identical frames, as human brain recognizes graphical information by ignoring objects that have already been seen [Ward, 2008].

To address the issues mentioned above, video summarization has been attracting more research and development efforts in recent years. Despite of variety of already released methods for video summarization, the

main problem they face is the gap between information retrieved from video and semantic description required for efficient summarization [Laganière, 2008]. Another great challenge is that frames are obtained under different exposure, lighting conditions, aperture, focus and focal length of camera. That is why several kinds of features should be considered at the same time to obtain complete description and find similar frames. To create short and simultaneously comprehensive overview of a video, meaningful visual features are described in section one, and similarity measures for frame partitions are provided in section two. Experimental results are given in the third section, and the last section presents our conclusions.

Feature Set for Frame Partitioning

Spatio-temporal features are usually calculated for salient points [Laganière, 2008], local areas (detected objects or regions of interest) [German, 2005], or the whole frame [Lin, 2013]. In order to find significant features, we should look for parameters picked out by humans for visual information interpretation. Color, texture and context features are three main components used by people for video understanding. These parameters are not considered separately from each other, as they are closely related [Haralick, 1973].

Color features are traditionally analyzed with histograms that depict frequency of occurrence of one or another color tone in the region of interest. Sometimes only one color channel with the most significant changes is taken for analysis, sometimes an average, maximum or minimum value from a local range of a histogram is used. Color features are often considered in a form of intensity which is the simplest way of image color presentation. It is computed as an average from red r, green g and blue b component of an image with RGB color scheme [Bezdek, 2005].

$$\text{intensity} = \frac{r + g + b}{3}. \tag{1}$$

Comparison of histograms H_1 and H_2 can be obtained from their intersection or by calculus of correlation $C(H_1, H_2)$ between them, where μ_1 and μ_2 are average values for H_1 and H_2 respectively [Lin, 2013]. The higher correlation is, the more alike two histograms are.

$$C(H_1, H_2) = \frac{\sum (H_1 - \mu_1)(H_2 - \mu_2)}{\sqrt{\sum (H_1 - \mu_1)^2} \sqrt{\sum (H_2 - \mu_2)^2}}. \tag{2}$$

Textural features contain information about spatial distribution of changes in color tone for the whole image or its local part. In order to characterize texture, one may use any of 28 textural features proposed in [Haralick, 1973]. Though, it is important to note that they highly correlate with each other. Despite these features were proposed in 1973, many contemporary scientists turn towards them [Schonfeld, 2010]. Very often statistic degree of randomness, called entropy, is chosen for texture analysis.

From the point of video analysis, entropy describes spatial relations between brightness of frame pixel pairs, where $p(i, j)$ is an element of normalized matrix that describes spatial distribution of color tones in a frame (or local region) [Haralick, 1973].

$$E = -\sum_i \sum_j p(i, j) \log_2 (p(i, j)). \tag{3}$$

High entropy indicates large scatter of pixel values, while low entropy says about pixel homogeneity (and details consequently). Thus, entropy shows how much details consist in a local region, for which entropy value has been calculated [German, 2005]. This hypothesis can be easily proved by taking the same image with different resolution. The higher resolution is, the more details are visualized and the higher entropy value is (though such entropy changes are not that significant for an image with the same content). Along with the mentioned above

methods for texture analysis, there are methods based on auto-regression, Markov chains, mathematical morphology, fractals, wavelets, etc. [Sonka, 2007].

Such features as brightness, estimation of object borders, area, shape (using geometric matching), absolute and relative location, density and speed of motion, trajectory and many others are often used. Motion density and speed are usually estimated by optical flow [Schonfeld, 2010]. Though, by application of optical flow we bulk significantly the computational procedure compared with any other feature set. Trajectories of object motion are calculated with the help of differential images which may not account direction of motion. To save information about motion direction, cumulative differential image should be used. Such kind of images also enables to save some other temporal properties of motion, motion of huge objects and slow motion [Sonka, 2007]. Another interesting approach consists in analysis of structural features. Object shape is flooded with water-filling algorithm, filling time and shape length are considered [Zhou, 2001].

Great success has been achieved during the past years at the level of image understanding. Despite this, many questions remain undecided in context analysis, and researchers continue working on this field [Sonka, 2007]. Contextual features are referred to high level features, they include information from graphical data blocks around the area of interest. Such image description assumes model development for each recognized object, identification of regions with potential object samples.

High level description can be performed with low level features, assuming their absolute or relative spatial location on images, or by application of artificial intelligence methods for their processing (such as fuzzy production rules, heuristics, cluster analysis, neural nets, different filters, etc.) [Depalov, 2006; Zhang, 200;, Schonfeld, 2010]. Examples of 44 systems, where such an approach is realized, are given at [Veltkamp, 2001]. Dominant colors, histogram analysis of separate regions and the whole frames, histogram correlation, coherent color vectors, mean colors are used as color features here. Border pixel statistics, local binary patterns, random field, elementary textural features, wavelet analysis and Fourier transformation are used as textural features in CBIR systems. Ellipses and bounding boxes, Fourier descriptors, elastic models, different curves and patterns are often used to define shape features.

Another approach to high level description lies in assignment of textual labels for different image classes by construction of semantic nets based on thesaurus. Textual label correspondence to the particular image class is defined by users who train a system. Such recognition algorithms search for similarity inside semantic networks and consider integrated visual features [Carneiro, 2007; Divakaran, 2009].

Although, due to inability of full-scale recognition implementation (similar to human mental activity), it is used to speak about mid level features that link semantic description with low level features [Boureau, 2010, Schonfeld, 2010]. For the purpose of frame partitioning we consider traditional color, textural and spatial features, taking into account relative location of regions and their regional properties, which give us a chance to obtain meaningful segments and extract key frames in future.

Matching of Frame Partitions

Frame partitioning is proposed to be performed with Voronoi diagrams. This method has been chosen because of several reasons. First of all, frame partitioning into real objects is not reasonable because of their tremendous changes in time. For now it is impossible to process all of them efficiently at the same time. And their changes may cause fault detection and object mess. Secondly, this partition technique requires less computational resources than real object segmentation and much less than motion analysis. And thirdly, Voronoi diagrams have not been used yet for the purpose of video summarization, thus, we want to develop a novel method and check its efficiency by comparing with existing ones.

Voronoi diagrams were first mentioned by R. Descartes in 1644. Later, in 1850, they were declared by P.G.L. Dirichlet, and further named after Russian mathematician G.F. Voronoi [Okabe, 2000]. To give formal definition of a Voronoi tessellation, let us denote $D=[a,b]\times[c,d]$, $a,b,c,d=const$ as a field of view. Let $\{p_1, p_2, ..., p_n\}$ be a finite set of generator points selected by Harris method that takes into account pixel intensity and relative location of regions. (Harris method [Sonka, 2007] has been chosen as one of the most frequently used with good performance and relative simplicity.) Voronoi diagram is a field of view partition $V=\{v(p_1)\cap D, v(p_2)\cap D, \ldots, v(p_n)\cap D\}$ into convex polygons, s.t.

$$v(p_i) = \{z \in R^2 : d(z,p_i) \le d(z,p_j)\ \forall i \ne j\} \tag{4}$$

where $d(\circ,\circ)$ is a planar Euclidean metric [Okabe, 2000].

To define a key frame, let $B_k(z), z=(x,y)\in D$ be the k-th frame from video sequence Φ (here and subsequently $k=1,2,\ldots,K$ is a discrete time). If $1\le i<j\le K$ and $B_i(z), B_j(z)\in\Phi$ then we shall use notation $S_l(i,j)=[B_i(z),B_j(z)]$, $l=1,2,\ldots$, $i,j\in L_l$, $\sum L_l = K$ for a scene that is a set of sequential frames obtained after temporal segmentation into meaningful segments, s.t. $\forall l\ S_l(i,j)\ne\varnothing$, $\Phi=\bigcup_{l\in L} S_l(i,j)$, $\forall l', l''\ S_{l'}(i,j)\cap S_{l''}(i,j)=\varnothing$. For a fixed l, define a key frame as an image $B_r^*(z)\in S_l(i,j)$ with property

$$r=\arg\min_{r\in L_l}\left(\sum_{t\in L_l, r\ne t}\rho(B_r(z),B_t(z))\right) \tag{5}$$

where $\rho(\circ,\circ)$ is a metric. After all the key frames are extracted, we obtain the set $\{B_l^*(z)\}$ of key frames for video stream Φ. In other words, we extract a frame (or several) per scene, and each key frame extracted is the most representative one for its scene (or subscene).

Incorporation of matching procedure has not been done yet for Voronoi tessellations, except by Yukio Sadahiro [Sadahiro, 2011]. He introduced different methods of visual and quantitative analysis, including χ^2, Kappa index and their extensions, area and perimeter of tessellations, their variance and standard deviation, spatial mean of their gravity centers, etc. His idea was to implement granularity density measure and hierarchy relationships (overlay, partial overlay and inclusion) to compare different Japanese administrative region division systems, though the areal methods are quite ambiguous for video processing application, as objects may be shot at different zoom. Different objects in images may possess the same area. Thus, video objects cannot be traced with properties primarily based on area. In our case different attributes are needed to be considered. For our purposes we used spatial, textural and color features which are among the main attributes used for CBIR and CBVR.

To match frame partitions, consider two frames $B'(z), B''(z)$ with generator points $\{p'_1, p'_2, ..., p'_n\}$ and $\{p''_1, p''_2, ..., p''_m\}$ respectively, then spatial dissimilarity of frames can be approximately represented by partition metric $\rho_1(V',V'')$ [Mashtalir, 2006]

$$\rho_1(V',V'')=\sum_{i=1}^{n}\sum_{j=1}^{m}\mathrm{card}(v(p'_i)\Delta v(p''_j))\,\mathrm{card}(v(p'_i)\cap v(p''_j)) \tag{6}$$

where $v(p'_i)\Delta v(p''_j)=(v(p'_i)\setminus v(p''_j))\cup(v(p''_j)\setminus v(p'_i))$ is symmetric difference that counts the number of elements on which $v(p'_i)$ and $v(p''_j)$ differ [Yianilos, 1991].

The above distance measure shows how two diagrams match each other in terms of regions. To take into account color and textural features, let us define two more metrics, $\rho_2(B'(z),B''(z))$ and $\rho_3(B'(z),B''(z))$

respectively which are defined in common regions of partitions. By measuring similarity in color and texture we observe changes between Voronoi regions of two frames being analyzed. Squared Euclidean distance has been used to incorporate more weight for distant color objects. Manhattan distance has been chosen for textural similarity measurement, as entropy values are calculated for the whole regions and they are presented by a single float value per region, while similarity in color is taken from each pixel present in both frames.

$$\rho_2(B'(z),B''(z))=\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{x_q}\sum_{y_u}{}_{(x_q,y_u)\in(v(p'_i)\cap v(p''_j))}(B'(x_q,y_u)-B''(x_q,y_u))^2\ ,$$
$$\rho_3(B'(z),B''(z))=\sum_{i=1}^{n}\sum_{j=1}^{m}{}_{(v(p'_i),v(p''_j))\supseteq(v(p'_i)\cap v(p''_j))}\left|E(v(p'_i))-E(v(p''_j)\right|\ . \tag{7}$$

where $B'(x_q,y_u)$ is intensity value for pixels in a region $(v(p'_i)\cap v(p''_j)$, and $E(v(p'_i))$ is entropy value in a region $v(p'_i)$.

Thus, we have got non-normalized estimates. For this reason we offer to normalize formulas (6) and (7) to obtain values ranging from 0 to 1. Conversion of the above metrics to bounded forms assumes application of a function, named range compander [Yianilos, 1991], s.t. its combination with a metric still gives a metric which satisfies non-negativity, reflexivity, symmetry and triangle inequality rules.

$$\rho'(B'(z),B''(z))=\frac{1}{1+\rho(B'(z),B''(z))} \tag{8}$$

As non-negative linear combination of metrics is still a metric, we may propose the following resulting metric:

$$\hat{\rho}(B'(z),B''(z))=\alpha_1\rho'_1+\alpha_2\rho'_2+\alpha_3\rho'_3\ ,\quad \sum_{\gamma=1}^{\gamma=3}\alpha_\gamma=1\ ,\ \alpha_\gamma\geq 0 \tag{9}$$

where $\hat{\rho}(B'(z),B''(z))$ shows similarity between frames, and α_γ shows the impact of each feature in use.

In order to extract frames with lowest level of proximity, we should compare consecutive frames pair-wise. Tessellation matching algorithm for key frame selection is described below.

1. **Determine homogeneity of video content.** Calculate texture variance (dispersion of entropy) throughout the video sequence, and set a threshold value according to the following rule:

$$Threshold=\begin{cases}\frac{1}{4},\ \frac{1}{K-1}\sum_{k=1}^{k=K}\left(E(B_k(z))-\frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z))\right)^2\to\infty\\ \frac{1}{2},\ \frac{1}{K-1}\sum_{k=1}^{k=K}\left(E(B_k(z))-\frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z))\right)^2\to\frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z))\\ \frac{3}{4},\ \frac{1}{K-1}\sum_{k=1}^{k=K}\left(E(B_k(z))-\frac{1}{K}\sum_{k=1}^{k=K}E(B_k(z))\right)^2\to 0\end{cases} \tag{10}$$

where $E(B_k(z))$ is entropy for k-th video frame, K is a total number of frames in the video sequence.

A threshold value should be set according to video content. For videos with heterogeneous content and variety of scenes (see fig. 2) this value should be less than $\frac{1}{4}$, not to extract too much frames. On the contrast, for videos with homogeneous content (see fig. 3) and small number of scenes (or even a single scene) this value should be increased up to $\frac{3}{4}$, to extract a bit more frames.

2. Take the first ($B_k(z)$) and the second ($B_{k+1}(z)$) frames for comparison. Set $k = 1$.

3. **Frame matching.** According to formula (9), compute $\hat{\rho}(B_k(z), B_{k+1}(z))$ for two frames. If $\hat{\rho}(B_k(z), B_{k+1}(z))$ is less than the predefined threshold value, then extract both frames $B_k(z)$ and $B_{k+1}(z)$ as key frames $B_r^*(z)$ and $B_{r+1}^*(z)$ and go to step 4, otherwise extract only $B_k(z)$ as a key frame $B_r^*(z)$ and go to step 5.

4. Set $B_k(z) = B_{k+1}(z)$, $B_{k+1}(z) = B_{k+2}(z)$ and go to step 6.

5. Leave $B_k(z) = B_k(z)$ and set $B_{k+1}(z) = B_{k+2}(z)$.

6. Repeat step 3, until $B_{k+1}(z) \le K$.

7. **Inter-scene key frame comparison.** Thus, we have obtained a key frame per scene: $\{B_1^*(z) \in S_1(i, j)\} \ldots \{B_l^*(z) \in S_l(i, j)\}$. Compare key frames pair-wise between scenes using frame matching procedure, defined in step 3. Second identical key frame is to be deleted. Thus, the resulting key frame sequence will be $\{\hat{B}_1^*(z), \ldots, \hat{B}_l^*(z)\}$.

Experimental Results

The proposed method has been tested on low resolution Trecvid video samples, several commercials of medium resolution and self-made high definition videos. Key frames extracted from Chinese commercial about Mercedes Benz C-Class automobile are shown in fig. 2. Examples of partitioning of frames (with homogeneous content and high definition) using Voronoi diagrams are provided in fig. 3.

Test results have been compared with existing summarization techniques based on clustering, curve simplification, and motion analysis. This comparison has shown good balance between high precision and recall for the proposed procedure. The estimation has been performed by 10 respondents who knew nothing about the name of frame extraction method they tested.

Figure 2. Key frames extracted from Chinese commercial

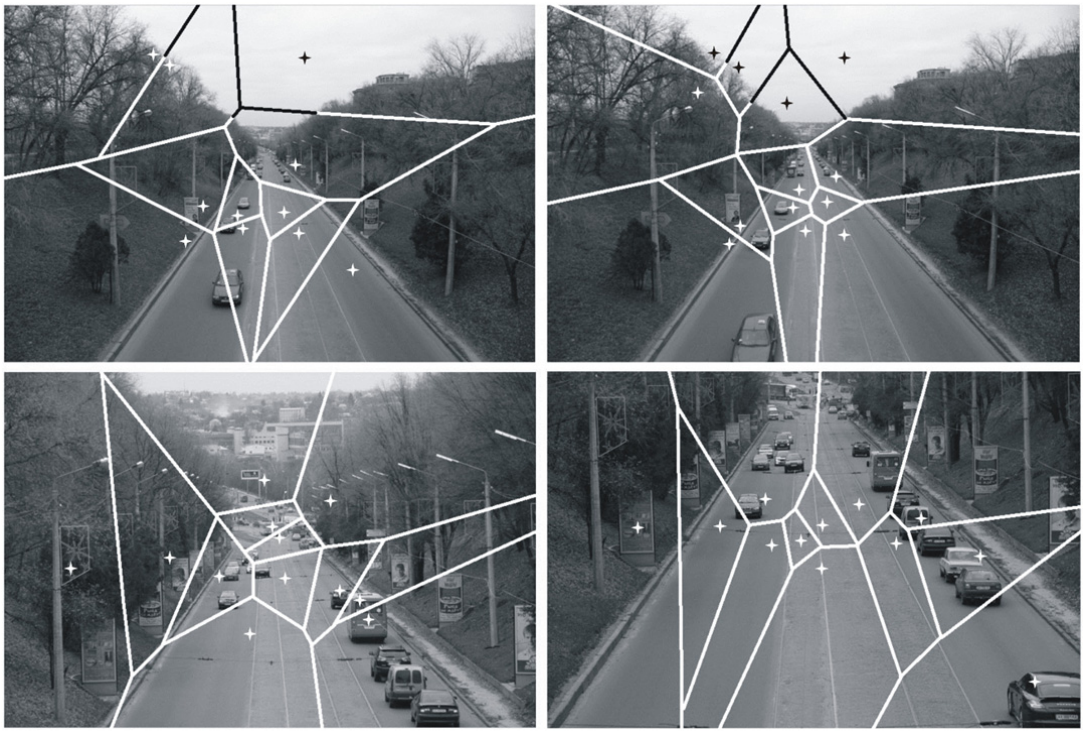


Figure 3. Examples of frame partitioning using Voronoi diagrams

Conclusion

By analyzing the difficulties faced by researchers during video summarization, we have come to the conclusion that the main problem lies in the gap between semantic content and low level frame presentation. To make an attempt of overcoming this gap, we have proposed a new method of key frame extraction based on Voronoi partitioning of frames, which assumes spatial features, color, texture, and relative location of regions.

The proposed method differs from existing ones in accuracy of results and computational uniqueness of matching procedure. The accuracy of results is reached due to generalized procedure of region processing. Existing algorithms reveal changes almost at each frame, though these changes may not be that important, while the proposed method returns only key frames with significant changes in content. Shape changes are dramatic at each frame, but Voronoi region is quite stable. It has been shown, that frames with identical content are partitioned in a similar manner with Voronoi diagrams.

The proposed method takes into account video content homogeneity by setting the appropriate threshold value before matching the frames. Key frames are compared with each other between video scenes, detected using the technique proposed in [Bodyanskiy, 2012]. Duplicate key frames are removed with the second pass of the algorithm.

Acknowledgements

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Bibliography

[Bezdek, 2005] J.C. Bezdek et al. Fuzzy models and algorithms for pattern recognition and image processing. New York: Springer, 2005, 776 p.

[Bodyanskiy, 2012] Y. Bodyanskiy et al. On-line video segmentation using methods of fault detection in multidimensional time sequences. In: International Journal of Electronic Commerce Studies, 2012, Vol. 3, No. 1, pp. 1-20.

[Boureau, 2010] Y.-L. Boureau et al. Learning Mid-Level Features For Recognition. In: Computer Vision and Pattern Recognition, 2010, pp. 2559-2566.

[Carneiro, 2007] G. Carneiro et al. Supervised Learning of Semantic Classes for Image Annotation and Retrieval. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 29, No. 3, 2007, pp. 394-410.

[Depalov, 2006] D. Depalov et al. Perceptual feature selection for semantic image classification. In: IEEE International Conference on Image Processing, Vol. 2, 2006, pp. 2921-2924.

[Divakaran, 2009] A. Divakaran. Multimedia Content Analysis: Theory and Applications (Signals and Communication Technology). New York: Springer, 2009, 390 p.

[German, 2005] A. German, M.R. Jenkin, Y. Lesp´erance. Entropy-based image merging. In: 2-nd Canadian Conference on Computer and Robot Vision, 2005, pp.81-86.

[Goldman, 2007] D.R. Goldman. Framework for video annotation, visualization, interaction: Doctoral Thesis. Washington, 2007. – 140 p.

[Haralick, 1973] R.M. Haralick, K. Shanmugam, I. Dinstein. Textural features for Image Classification. In: IEEE transactions on systems, man and cybernetics, Vol. 3, No. 6, 1973, pp. 610-621.

[Laganière, 2008] R. Laganière et al. Video Summarization from Spatio-Temporal Features. In: 2-nd ACM TRECVid Video Summarization Workshop, 2008, pp. 144-148.

[Lin, 2013] G.-S. Lin, J.-F. Chang. Detection of frame duplication forgery in videos based on spatial and temporal analysis. In: International Journal of Pattern Recognition and Artificial Intelligence, Vol. 26, No. 7, 2013.

[Mashtalir, 2006] V. Mashtalir et al. A novel metric on partitions for image segmentation. In: IEEE International Conference on Video and Signal Based Surveillance, 2006.

[Okabe, 2000] A. Okabe et al. Spatial tessellations: Concepts and applications of Voronoi diagrams. – 2-nd ed. – Chichester: Wiley, 2000, 671 p.

[Rubin, 2005] M. Rubin. Droidmaker: George Lucas and the digital revolution. – Gainesville: Triad Publishing, 2005, 518 p.

[Sadahiro, 2011] Y. Sadahiro. Analysis of the relationship among spatial tessellations. In: Journal of Geographical Systems, Vol. 13, No. 4, 2011, pp. 373-391.

[Schonfeld, 2010] D. Schonfeld et al. Video Search and Mining. In: Studies in Computational Intelligence, Vol. 287, Berlin: Springer, 2010, 388 p.

[Sonka, 2007] M. Sonka, V. Hlavac, R. Boyle. Image Processing, Analysis, and Machine Vision, International Student Edition. – 3 ed. – Toronto: Thomson, 2007, 850 p.

[Veltkamp, 2001] R.C. Veltkamp, H. Burkhardt, H.-P. Kriegel. State-of-the-Art in Content-Based Image and Video Retrieval (Computational Imaging and Vision). Netherlands: Kluwer Academic Publishers, 2001, 343 p.

[Ward, 2008] K. Ward. Augenblick: The Concept of the 'Decisive Moment' in 19th and 20th Century Western Philosophy. Aldershot: Ashgate, 2008, 192 p.

[Yianilos, 1991] P.N. Yianilos. Normalized forms for two common metrics. In: NEC Research Institute, Report 91-082-9027-1, 1991, Revision 7/7/2002. Cambridge: Cambridge University Press, 1991, 7 p.

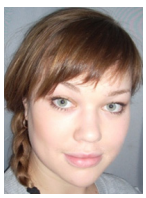
[Zhang, 2008] D. Zhang, Y. Liu, J. Hou. Digital Image Retrieval Using Intermediate Semantic Features and Multistep Search. In: Digital Image Computing: Techniques and Applications, 2008, pp. 513-518.

[Zhou, 2001] X.S. Zhou, T.S. Huang. Edge-Based Structural Features for Content-Based Image Retrieval. In: Pattern Recognition Letters, Vol. 22, No. 5, 2001, pp. 457-468.

Authors' Information

Olena Mikhnova *– PhD student of Informatics department in Kharkiv National University of Radio Electronics, P.O. Box: 61166, Ukraine, Kharkiv, Lenina av., 14; e-mail: elena_mikhnova@ukr.net*

Major Fields of Scientific Research: Image and video recognition, Data mining



Nataliia Vlasenko *– PhD student of Informatics department in Kharkiv National University of Radio Electronics, P.O. Box: 61166, Ukraine, Kharkiv, Lenina av., 14; e-mail: gorohovatskaja@gmail.com*

Major Fields of Scientific Research: Image recognition and analysis

О ПРИБЛИЖЕННОМ РЕШЕНИИ ЗАДАЧ ВОССТАНОВЛЕНИЯ ЗАВИСИМОСТЕЙ С ПОМОЩЬЮ АЛГОРИТМОВ РАСПОЗНАВАНИЯ

Владимир Рязанов, Антон Щичко

***Abstract**: Предлагается подход к восстановлению зависимостей, основанный на решении задач распознавания (классификации с учителем). Вычисление значения зависимой величины сводится к распознаванию интервала, которому она принадлежит. По данным обучения поставлена задача поиска оптимального разбиения области допустимых значений зависимой величины. Задача сформулирована в виде дискретной оптимизационной задачи. Вычисление значения оптимизируемой функции требует решения большого числа обучения распознаванию в режиме скользящего контроля. Для нескольких моделей распознавания типа вычисления оценок получены формулы быстрого переобучения при переходе от одной задачи распознавания к соседней. Приводятся результаты применения созданной модели и алгоритма восстановления значения зависимой величины для решения практических задач.*

***Keywords**: восстановление зависимости, регрессия, классификация, алгоритм распознавания, прецедент, кусочно-постоянная функция, признак*

Введение

Многие задачи анализа прецедентных данных часто задаются в следующем стандартном виде. Дана выборка $\{z_i, \mathbf{x}_i\}, i = 1,2,...,m$, где $\mathbf{x}_i = (x_{i1}, x_{i2},..., x_{in})$ - признаковое описание объекта, $z_i \in R$, $x_{ij} \in M_j$ (M_j - известное множество допустимых значений признака № j). Вектор значений $\mathbf{x}_i$ задает исходные характеристики объекта, которые можно вычислить или измерить. Поэтому мы будем называть его вектором значений независимых параметров. Предполагается, что величина z_i , известная для объектов обучающей выборки, является скрытой от наблюдения главной характеристикой объекта, которая может быть вычислена по вектору $\mathbf{x}_i$, т.е. $z_i = f(\mathbf{x}_i)$. Будем ее называть зависимой величиной. Требуется для произвольного нового $\mathbf{x} = (x_1, x_2,..., x_n)$, $x_j \in M_j$, вычислить $z = f(\mathbf{x})$, $z \in R$, используя данные обучающей выборки. В статистической постановке данная задача известна как задача восстановления регрессии - функции условного математического ожидания, при предположении существования условной плотности $p(z \mid \mathbf{x})$. В данной статье не будут использоваться какие-либо вероятностные модели и предположения.

В настоящее время существуют различные параметрические и непараметрические подходы к восстановлению зависимостей при $x_i \in R, i = 1,2,..., n$.

Параметрический подход [Дрейпер, Смит, 2007] предполагает наличие функциональной зависимости от некоторых параметров $\boldsymbol{\omega}$, причем вид зависимости известен:

- линейная зависимость - $f(\boldsymbol{\omega}, \mathbf{x}) = \sum_{j=1}^{n} \omega_j x_j + \omega_0$;

- полиномиальная зависимость степени γ - $f(\boldsymbol{\omega}, \mathbf{x}) = \sum_{p_1=0}^{\gamma_1} \ldots \sum_{p_n=0}^{\gamma_n} \omega_{p_1 \ldots p_n} x_1^{p_1} \ldots x_n^{p_n}$, $\gamma = \sum_{i=1}^{n} \gamma_i$;
- криволинейная зависимость - $f(\boldsymbol{\omega}, \mathbf{x}) = \sum_{j=1}^{k} \omega_j \phi_j(x_1, \ldots, x_n) + \omega_0$, $\phi_1, \ldots, \phi_k$ - преобразования $R^n \to R$;
- логистическая зависимость: $f(\boldsymbol{\omega}, \mathbf{x}) = \frac{1}{1 + \exp(-z)}$, $z = \sum_{j=1}^{n} \omega_j x_j + \omega_0$.

В непараметрическом подходе [Хардле, 1993] характеристика z для $\mathbf{x}$ определяется как $z = \frac{\sum_{i=1}^{m} \omega_i(\mathbf{x}) z_i}{\sum_{i=1}^{m} \omega_i(\mathbf{x})}$, где $\omega_i(\mathbf{x}) = K\left(\frac{\rho(\mathbf{x}, \mathbf{x}_i)}{h}\right)$, $i = 1,2,\ldots,m$, K - ядерная функция, h - ширина окна.

Широко известны методы регрессии основанные на опорных векторах [Collobert, Bengio, 2001], которые могут рассматриваться как разновидность криволинейных регрессий. В работе [Jing-Rung Yu, Gwo-Hshiung Tzeng, Han-Lin Li 2001] был предложен общий метод построения нечеткой кусочной регрессии, где точки разбиения для значений зависимой величины вычисляются одновременно как решения смешанной задачи математического программирования. Существуют и другие близкие по сути подходы.

Отметим главные ограничения данных подходов. Параметрические подходы требуют априорного знания аналитического вида функций. Наличие разнотипных признаков (вещественных, номинальных, бинарных, порядковых, и т.п.) требует привлечения дополнительных средств описания объектов в единой шкале. Непараметрические методы широко используют частотные оценки, функции расстояний, что может быть весьма приближенным и практически затруднительным для выборок малой длины, при большом числе независимых параметров, различной их информативности и разнотипности, при наличии шумовых признаков.

В настоящей статье предлагается подход, не использующий априорные вероятностные предположения и основанный на теории распознавания (классификации с учителем). Задача решается следующим образом. Каждое разбиение отрезка $[a,b]$, $a = \min_{i=1,2,\ldots,m} z_i$, $b = \max_{i=1,2,\ldots,m} z_i$ на конечное число отрезков порождает некоторую задачу распознавания по прецедентам и определяется значениями некоторого вектора параметров $\mathbf{y}$ с конечным числом возможных значений. Для стандартной задачи распознавания по прецедентам $\mathbf{Z}(\mathbf{y})$ решается задача обучения и находится алгоритм классификации $A(\mathbf{y})$, качество которого оценивается функцией $F(A(\mathbf{y}))$ в режиме скользящего контроля (leave-one-out procedure). Нахождение оптимального числа точек разбиения и самих точек сводится в минимизации $F(A(\mathbf{y}))$. С каждым классом связывается некоторое значение зависимой величины, оцениваемое по объектам обучения данного класса (выборочное среднее, среднее от пары точек, и т.д.). Вычисление функции $F(A(\mathbf{y}))$ в режиме скользящего контроля требует многократного обучения для соседних задач. Для некоторых алгоритмов типа вычисления оценок получены эффективные формулы переобучения алгоритмов при переходе от одной стандартной задачи распознавания к соседней ей. Вычисление

значения зависимой величины для некоторого $\mathbf{x}$ сводится к распознаванию класса объекта относительно найденного оптимального разбиения.

Далее, без ограничения общности, будем считать, что значения z_i различны, а объекты обучающей выборки упорядочены по возрастанию значений z_i, т.е. $z_i < z_{i+1}, i = 1,2,...,m-1$.

Постановка задачи восстановления зависимостей в классе кусочно-постоянных функций

Пусть фиксирован некоторый вектор $\mathbf{y} = (y_1,...,y_{l-1})$, $y_1 < y_2 < ... < y_{l-1}$, $z_2 \le y_i \le z_{m-2}, i = 1,2,...,l-1,$ $y_i < y_{i+1}, i = 1,2,...,l-2$, $y_i \in \{z_j \mid j = 1,2,...,m\}$. Набор $\mathbf{y}$ определяет разбиение вещественной оси на l множеств $I_1 = (-\infty, y_1], I_2 = (y_1, y_2],..., I_l = (y_{l-1}, +\infty)$ и разбиение множества допустимых признаковых описаний объектов на l классов $K_j = \{\mathbf{x} : z = f(\mathbf{x}) \in I_j\}, j = 1,2,...,l$. Множества K_j задаются в виде $\widetilde{K}_j = \{\mathbf{x}_i : z_i \in I_j, i = 1,2,...,m\}, j = 1,2,...,l$.

Пусть фиксирован некоторый алгоритм классификации $A^{\mathbf{y}}$ относительно классов $K_j, j = 1,2,...,l$, заданных вектором $\mathbf{y}$. Через $\widetilde{A}_i^{\mathbf{y}}$ обозначим алгоритм распознавания, соответствующий вектору $\mathbf{y}$, после удаления из обучающей выборки объекта $\mathbf{x}_i$. Определим функцию $f(\mathbf{x} \mid A^{\mathbf{y}})$ следующим образом:

$$f(\mathbf{x} \mid A^{\mathbf{y}}) = \begin{cases} (y_1 + z_1)/2, & \mathbf{x} \xrightarrow{A^{\mathbf{y}}} K_1, \\ (y_j + y_{j-1})/2, & \mathbf{x} \xrightarrow{A^{\mathbf{y}}} K_j, \quad j = 2,3,...,l-1, \\ (z_m + y_{l-1})/2, & \mathbf{x} \xrightarrow{A^{\mathbf{y}}} K_l, \end{cases} \qquad (1)$$

если объект $\mathbf{x}$ отнесен алгоритмом $A^{\mathbf{y}}$ в некоторый класс. Данный результат классификации будем обозначать как $A^{\mathbf{y}}(\mathbf{x}) = j$. Соответственно, $\widetilde{A}_i^{\mathbf{y}}(\mathbf{x}_i) = j$ будет обозначать результат распознавания (номер класса) алгоритмом $\widetilde{A}_i^{\mathbf{y}}$ объекта $\mathbf{x}_i$.

Будем использовать обозначение $|I_t| = |\{z_i : z_i \in I_t\}|, t = 1,2,...l$.

Рассмотрим следующую оптимизационную задачу.

$$F(y_1, y_2,..., y_{l-1}) = \sum_{z_i : z_i \le y_1} \left| z_i - f(\mathbf{x}_i \mid \widetilde{A}_i^{\mathbf{y}}) \right| + \sum_{i=1}^{l-2} \sum_{z_i : y_i < z_i \le y_{i+1}} \left| z_i - f(\mathbf{x}_i \mid \widetilde{A}_i^{\mathbf{y}}) \right| + \sum_{z_i : y_{l-1} \le z_i} \left| z_i - f(\mathbf{x}_i \mid \widetilde{A}_i^{\mathbf{y}}) \right| \to \min,$$

$$z_3 \le y_i \le z_{m-2}, i = 1,2,...,l-1, \; y_i < y_{i+1}, i = 1,2,...,l-2, \; |I_1| \ge 3, \; |I_t| \ge 2, t = 2,3,...l. \qquad (2)$$

Ограничения на $|I_t| = |\{z_i : z_i \in I_t\}|, t = 1,2,...l$, связаны с видом функции (1), поскольку значение зависимой величины вычисляется по двум точкам выборки, и процедурой скользящего контроля.

Оптимальное решение $y_1, y_2,..., y_{l-1}$ и определяет точки разбиения значений зависимой величины и оптимальную стандартную задачу распознавания.

После нахождения оптимальных значений $y_1, y_2,..., y_{l-1}$, вычисление значений зависимой величины осуществляется следующим образом.

1. Решается задача классификации объекта $\mathbf{x}$ алгоритмом $A^{\mathbf{y}}$.
2. Вычисление $\mathbf{z} = f(\mathbf{x}) \equiv f(\mathbf{x}|A^{\mathbf{y}})$ проводится по формуле (1).

Отметим, что при вычислении по формуле (1) значения зависимой величины мы использовали среднее по граничным точкам. Здесь могут быть использованы и другие способы ее оценки (выборочное среднее по классу, медиана, и т.п.). При классификации $\mathbf{x}$ мы считаем, что результат «отказ от классификации» не существует. В данном подходе может быть использована произвольная модель распознавания. Основная вычислительная проблема здесь связана с вычислением $F(y_1, y_2, ..., y_{l-1})$ в режиме скользящего контроля, когда требуется быстрое переобучение алгоритма при исключении некоторого объекта из обучающей выборки. Данное переобучение осуществляется быстро в некоторых моделях распознавания типа вычисления оценок. Далее мы рассмотрим общую схему оптимизации и используемые алгоритмы распознавания. Отметим, что вычисляемая в итоге функция является кусочно-постоянной.

Алгоритм поиска оптимальной кусочно-постоянной функции по обучающей выборке

Рассмотрим задачу (2), где функция $f(\mathbf{x})$ (и соответствующий алгоритм распознавания A), задаются текущими значениями параметров $y_1, y_2, ..., y_{l-1}$. Объекты обучения упорядочены по возрастанию z_i и имеет место $y_t = z_{i_t}$. Рассматривалась схема локальной оптимизации $F(y_1, y_2, ..., y_{l-1})$. Точки $\mathbf{y}^* = (y_1^*, y_2^*, ..., y_{l-1}^*)$, $y_j^* \in \{z_{i_j-1}, z_{i_j+1}\}$, $y_t^* = y_t, t \neq j$, $j = 1,2,...,l-1$, назовем соседними для $\mathbf{y} = (y_1, y_2, ..., y_{l-1})$. Начиная с произвольного допустимого $\mathbf{y}^{(0)} = (y_1^{(0)}, y_2^{(0)}, ..., y_{l-1}^{(0)})$ осуществляем просмотр всех не более чем $2(l-1)$ соседних допустимых (т.е. удовлетворяющих ограничениям (2)) точек. Обозначим через $\mathbf{y}' = (y_1', y_2', ..., y_{l-1}')$ произвольную соседнюю допустимую точку для $\mathbf{y}^{(0)} = (y_1^{(0)}, y_2^{(0)}, ..., y_{l-1}^{(0)})$. В качестве точки минимума $F(y_1, y_2, ..., y_{l-1})$ в окрестности $\mathbf{y}^{(0)}$ принимаем произвольную соседнюю допустимую точку $\mathbf{y}^{(1)} = (y_1^{(1)}, y_2^{(1)}, ..., y_{l-1}^{(1)})$, для которой $F(y_1^{(1)}, y_2^{(1)}, ..., y_{l-1}^{(1)}) \leq F(y_1', y_2', ..., y_{l-1}')$, $F(y_1^{(1)}, y_2^{(1)}, ..., y_{l-1}^{(1)}) < F(y_1^{(0)}, y_2^{(0)}, ..., y_{l-1}^{(0)})$. Осуществляется переход в точку $\mathbf{y}^{(1)}$ и процесс повторяется. Алгоритм локальной оптимизации заканчивает работу на шаге t, если выполняется условие $F(y_1^{(t-1)}, y_2^{(t-1)}, ..., y_{l-1}^{(t-1)}) \leq F(y_1', y_2', ..., y_{l-1}')$ для всех соседних вектору $(y_1^{(t-1)}, y_2^{(t-1)}, ..., y_{l-1}^{(t-1)})$ допустимых точек $(y_1', y_2', ..., y_{l-1}')$. Конечность локальной оптимизации следует из конечности допустимых значений $F(y_1, y_2, ..., y_{l-1})$. Отметим, что алгоритм динамического программирования [Михалевич, 1965] в здесь не применим, поскольку слагаемые в (2) зависят не только от значений одного-двух переменных y_i.

Модификация алгоритма вычисления оценок (ABO). Восстановление зависимостей с использованием модели распознавания ABO.

Опишем принцип работы алгоритмов вычисления оценок [Журавлев, Никифоров, 1971]. В выше приведенных обозначениях $z_i, z \in \{1,2,...,l\}$. Пусть некоторое множество X допустимых объектов $\mathbf{x}$

имеет вид $X = \bigcup_{j=1}^{l} K_j, K_\nu \cap K_\mu = \emptyset,\ \nu \neq \mu$. Дана обучающая выборка $\{z_t, \mathbf{x}_t, t = 1,2,...,m\}$, где $z_t = j$, если $\mathbf{x}_t \in K_j$. Для простоты считаем, что $\mathbf{x} = (x_1, x_2, ..., x_n) \in R^n$, а обучающая выборка содержит представителей всех классов. Обозначим $\widetilde{K}_j = \{\mathbf{x}_i : \mathbf{x}_i \in K_j, i = 1,2,...,m\}$. Пусть фиксирована система опорных множеств $\Omega_A = \{\Omega\}$, $\Omega \subseteq \{1,2,...,n\}$ алгоритма A. Опорное множество Ω задает некоторое подмножество признаков. Близость распознаваемого объекта $\mathbf{x}$ к некоторому объекту обучения $\mathbf{x}_t$ по опорному множеству Ω определяется как

$$\mathrm{B}_\Omega(\mathbf{x}, \mathbf{x}_t) = \begin{cases} 1, & |x_i - x_{ti}| \le \varepsilon_i, \forall i \in \Omega, \\ 0, & \text{иначе.} \end{cases} \quad (3)$$

В [Журавлев, Никифоров, 1971] присутствуют числовые параметры $\varepsilon_i, i = 1,2,...,n$, задаваемые пользователем или вычисляемые как, например, $\varepsilon_i = \frac{2}{m(m-1)} \sum_{\alpha,\beta=1,\alpha<\beta}^{m} |x_{\alpha i} - x_{\beta i}|$. Для объекта $\mathbf{x}$ вычисляется оценка $\Gamma_j(\mathbf{x})$ за класс $K_j, j = 1,2,...,l$:

$$\Gamma_j(\mathbf{x}) = \frac{1}{|\widetilde{K}_j|} \sum_{\mathbf{x}_i \in \widetilde{K}_j} \sum_{\Omega \in \Omega_A} B_\Omega(\mathbf{x}, \mathbf{x}_i). \quad (4)$$

Оценка $\Gamma_j(\mathbf{x})$ характеризует эвристическую степень близости объекта $\mathbf{x}$ к классу K_j. Далее применяется решающее правило в пространстве оценок: объект $\mathbf{x}$ относится алгоритмом A в класс K_j, если $\Gamma_j(\mathbf{x}) > \Gamma_i(\mathbf{x}), \forall i \neq j$. В противном случае выбор класса происходит из классов с максимальными оценками случайно. Обычно используют в качестве системы опорных множеств $\Omega_A = \{\Omega : |\Omega| = k\}$, где $0 \le k \le n$, k - целое, либо Ω_A есть система всех подмножеств множества $\{1,2,...,n\}$. Параметр k является внешним, мы использовали $k = \left[\frac{n}{3}\right]$. В [Журавлев, Никифоров, 1971] доказано, что при первом способе выбора $\Gamma_j(\mathbf{x}) = \frac{1}{|\widetilde{K}_j|} \sum_{\mathbf{x}_i \in \widetilde{K}_j} \mathrm{C}^k_{d(\mathbf{x},\mathbf{x}_i)}$, а во втором случае $\Gamma_j(\mathbf{x}) = \frac{1}{|\widetilde{K}_j|} \sum_{\mathbf{x}_i \in \widetilde{K}_j} (2^{d(\mathbf{x},\mathbf{x}_i)} - 1)$. Здесь $d(\mathbf{x}, \mathbf{x}_i) = \left|\{j : |x_j - x_{ij}| \le \varepsilon_j, j = 1,2,...,n\}\right|$.

В настоящей работе использовалась следующая модификация функции близости (3) и формулы (4). Пусть $\mathbf{x}_\alpha, \mathbf{x}_\beta \in \widetilde{K}_j$, тогда определим функцию близости $\widetilde{\mathrm{B}}_\Omega(\mathbf{x}_\alpha, \mathbf{x}, \mathbf{x}_\beta)$ объекта $\mathbf{x}$ к паре $\mathbf{x}_\alpha, \mathbf{x}_\beta$, и его оценку $\widetilde{\Gamma}_j(\mathbf{x})$ за класс K_j следующими выражениями.

$$\widetilde{\mathrm{B}}_\Omega(\mathbf{x}_\alpha, \mathbf{x}, \mathbf{x}_\beta) = \begin{cases} 1, & (x_{\alpha i} \le x_i \le x_{\beta i}) \vee (x_{\beta i} \le x_i \le x_{\alpha i}), \forall i \in \Omega, \\ 0, & \text{иначе.} \end{cases} \quad (5)$$

$$\widetilde{\Gamma}_j(\mathbf{x}) = \frac{2}{|\widetilde{K}_j|(|\widetilde{K}_j| - 1)} \sum_{\mathbf{x}_\alpha, \mathbf{x}_\beta \in \widetilde{K}_j, \alpha<\beta} \left(\sum_{\Omega \in \Omega_A} \widetilde{\mathrm{B}}_\Omega(\mathbf{x}_\alpha, \mathbf{x}, \mathbf{x}_\beta) \right). \quad (6)$$

Можно показать, что здесь также справедливы аналогичные эффективные формулы для вычисления оценок:

$\widetilde{\Gamma}_j(\mathbf{x}) = \frac{2}{\left|\widetilde{K}_j\right|\left(\left|\widetilde{K}_j\right|-1\right)} \sum_{\mathbf{x}_\alpha,\mathbf{x}_\beta \in \widetilde{K}_j, \alpha<\beta} \mathrm{C}^k_{d(\mathbf{x}_\alpha,\mathbf{x},\mathbf{x}_\beta)}$, $\widetilde{\Gamma}_j(\mathbf{x}) = \frac{2}{\left|\widetilde{K}_j\right|\left(\left|\widetilde{K}_j\right|-1\right)} \sum_{\mathbf{x}_\alpha,\mathbf{x}_\beta \in \widetilde{K}_j, \alpha<\beta} (2^{d(\mathbf{x}_\alpha,\mathbf{x},\mathbf{x}_\beta)} - 1)$, где

$d(\mathbf{x}_\alpha,\mathbf{x},\mathbf{x}_\beta) = \left|\{j : (x_{\alpha i} \le x_i \le x_{\beta i}) \vee (x_{\beta i} \le x_i \le x_{\alpha i}), j = 1,2,...,n\}\right|$. После вычисления оценок $\widetilde{\Gamma}_j(\mathbf{x}), j = 1,2,...,l,$ используется приведенное ранее решающее правило. Отметим, что в данном случае не используется метрика в R и параметры $\varepsilon_i, i = 1,2,...,n$. Признаки могут быть порядковыми.

Настоящий алгоритм не содержит параметров, требующих настройки при обучении.

Обозначим $m_j = \left|\widetilde{K}_j\right|, j = 1,2,...,l$. Рассмотрим задачу оптимизации (2), когда в качестве базовой модели распознавания используется модифицированная модель вычисления оценок.

При решении задачи оптимизации (2) оценки $\widetilde{\Gamma}_j(\mathbf{x})$ легко пересчитываются в режиме скользящего контроля.

Действительно, вычислим матрицы, $\mathbf{D}^1 = \left\|D^1_{\alpha\gamma\beta}\right\|_{m\times m\times m}$, $D^1_{\alpha\gamma\beta} = \mathrm{C}^k_{d(\mathbf{x}_\alpha,\mathbf{x}_\gamma,\mathbf{x}_\beta)}$, $\mathbf{D}^2 = \left\|D^2_{\alpha\gamma\beta}\right\|_{m\times m\times m}$, $D^2_{\alpha\gamma\beta} = 2^{d(\mathbf{x}_\alpha,\mathbf{x}_\gamma,\mathbf{x}_\beta)} - 1$. Пусть $\mathbf{x}_t$ - произвольный объект выборки, имеется текущее разбиение на классы $K_1, K_2, ..., K_l$ и $\mathbf{x}_t \in K_i$. Тогда $\widetilde{\Gamma}_i(\mathbf{x}_t) = \frac{2}{(m_i-1)(m_i-2)} \sum_{\substack{\alpha<\beta:\alpha,\beta\ne t \\ \mathbf{x}_\alpha,\mathbf{x}_\beta \in K_i,}} D^h_{\alpha t\beta}$,

$\widetilde{\Gamma}_j(\mathbf{x}_t) = \frac{2}{m_j(m_j-1)} \sum_{\substack{\alpha<\beta: \\ \mathbf{x}_\alpha,\mathbf{x}_\beta \in K_j,}} D^h_{\alpha t\beta}, j \ne i$. Здесь $h \in \{1,2\}$ и для простоты записи мы его опускаем далее в выражениях $\widetilde{\Gamma}_j(\mathbf{x}_t)$. Рассмотрим пересчет оценок $\widetilde{\Gamma}_j(\mathbf{x}_t), j = 1,2,...,l$, при пересчете функционала в соседней точке произвольного шага алгоритма оптимизации. При этом граница между некоторой парой классов меняется в результате переноса некоторого объекта $\mathbf{x}_\tau$ из одного класса в соседний класс. Обозначим «новые» классы $K^*_1, K^*_2, ..., K^*_l$, а оценки для $\mathbf{x}_t$ через $\widetilde{\Gamma}^*_j(\mathbf{x}_t), j = 1,2,...,l$.

Возможны следующие четыре варианта:

1. $K^*_i = K_i \cup \{\mathbf{x}_\tau\}$, $\mathbf{x}_\tau \in K_u, u \ne i$,

$K^*_u = K_u \setminus \{\mathbf{x}_\tau\},$

$K^*_j = K_j, j \ne i, u$.

2. $K^*_u = K_u \setminus \{\mathbf{x}_\tau\},$

$K^*_v = K_v \cup \{\mathbf{x}_\tau\},$

$K^*_j = K_j, j \ne u, v, \quad u, v \ne i$.

3. $K_i^* = K_i \setminus \{\mathbf{x}_\tau\}, \tau \neq t,$

$K_u^* = K_u \cup \{\mathbf{x}_\tau\},$

$K_j^* = K_j, j \neq i, u.$

4. $K_i^* = K_i \setminus \{\mathbf{x}_t\},$

$K_u^* = K_u \cup \{\mathbf{x}_t\},$

$K_j^* = K_j, j \neq i, u.$

Тогда оценки $\widetilde{\Gamma}_j^*(\mathbf{x}_t), j = 1,2,...,l$, пересчитываются следующим образом:

1. $$\widetilde{\Gamma}_i^*(\mathbf{x}_t) = \frac{2}{m_i(m_i-1)}\left(\frac{(m_i-1)(m_i-2)}{2}\widetilde{\Gamma}_i(\mathbf{x}_t) + \sum_{\substack{\alpha:\alpha\neq t \\ \mathbf{x}_\alpha\in K_i, \\ \mathbf{x}_\tau\in K_u, u\neq i}} D^h_{\alpha t\tau}\right),$$

$$\widetilde{\Gamma}_u^*(\mathbf{x}_t) = \frac{2}{(m_u-1)(m_u-2)}\left(\frac{m_u(m_u-1)}{2}\widetilde{\Gamma}_u(\mathbf{x}_t) - \sum_{\substack{\alpha:\mathbf{x}_\alpha\in K_u, \\ \mathbf{x}_\tau\in K_u, u\neq i}} D^h_{\alpha t\tau}\right),$$

$\widetilde{\Gamma}_j^*(\mathbf{x}_t) = \widetilde{\Gamma}_j(\mathbf{x}_t), j \neq i, u.$

2. $$\widetilde{\Gamma}_u^*(\mathbf{x}_t) = \frac{2}{(m_u-1)(m_u-2)}\left(\frac{m_u(m_u-1)}{2}\widetilde{\Gamma}_u(\mathbf{x}_t) - \sum_{\substack{\alpha:\mathbf{x}_\alpha\in K_u, \\ \mathbf{x}_\tau\in K_u, u\neq i}} D^h_{\alpha t\tau}\right),$$

$$\widetilde{\Gamma}_v^*(\mathbf{x}_t) = \frac{2}{(m_v+1)m_v}\left(\frac{m_v(m_v-1)}{2}\widetilde{\Gamma}_v(\mathbf{x}_t) + \sum_{\substack{\alpha:\mathbf{x}_\alpha\in K_v, \\ \mathbf{x}_\tau\in K_u, u\neq i}} D^h_{\alpha t\tau}\right) u, v \neq i,$$

$\widetilde{\Gamma}_j^*(\mathbf{x}_t) = \widetilde{\Gamma}_j(\mathbf{x}_t), j \neq u, v.$

3. $$\widetilde{\Gamma}_i^*(\mathbf{x}_t) = \frac{2}{(m_i-2)(m_i-3)}\left(\frac{(m_i-1)(m_i-2)}{2}\widetilde{\Gamma}_i(\mathbf{x}_t) - \sum_{\substack{\alpha:\mathbf{x}_\alpha\in K_i, \\ \mathbf{x}_\tau\in K_i, \tau\neq t}} D^h_{\alpha t\tau}\right),$$

$$\widetilde{\Gamma}_u^*(\mathbf{x}_t) = \frac{2}{(m_u+1)m_u}\left(\frac{m_u(m_u-1)}{2}\widetilde{\Gamma}_u(\mathbf{x}_t) + \sum_{\substack{\alpha:\mathbf{x}_\alpha\in K_u, \\ \mathbf{x}_\tau\in K_i, \tau\neq t}} D^h_{\alpha t\tau}\right),$$

$\widetilde{\Gamma}_j^*(\mathbf{x}_t) = \widetilde{\Gamma}_j(\mathbf{x}_t), j \neq i, u.$

4. $\widetilde{\Gamma}_i^*(\mathbf{x}_t) = \widetilde{\Gamma}_i(\mathbf{x}_t),$

$$\widetilde{\Gamma}_u^*(\mathbf{x}_t)=\frac{2}{(m_u+1)m_u}\left(\frac{m_u(m_u-1)}{2}\widetilde{\Gamma}_u(\mathbf{x}_t)+\sum_{\substack{\alpha<\beta:\\ \mathbf{x}_\alpha,\mathbf{x}_\beta\in K_u,}} D^h_{\alpha\beta}\right), u\neq i,$$

$$\widetilde{\Gamma}_j^*(\mathbf{x}_t)=\widetilde{\Gamma}_j(\mathbf{x}_t),\ j\neq i,u.$$

Таким образом, вычисление функционала в соседней точке общего алгоритма оптимизации осуществляется эффективно.

Восстановление зависимостей с использованием модели распознавания, основанной на голосовании по системам логических закономерностей классов

Алгоритмы распознавания, основанные на голосовании по системам логических закономерностей классов, описаны в [Рязанов, 2007; Ковшов, Моисеев, Рязанов, 2008] и работают следующим образом. Рассматривается множество элементарных предикатов, зависящих от числовых параметров $c_j^1, c_j^2, j=1,2,...,n:\ P_j^{c_j^1,c_j^2}(x)=\begin{cases}1, & c_j^1\le x\le c_j^2,\\ 0, & \text{иначе}\end{cases}$. Пусть $\Omega\subseteq\{1,2,...,n\}$.

Определение [Рязанов, 2007] . Предикат $P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x})=\&_{j\in\Omega}P_j^{c_j^1,c_j^2}(x_j)$ называется логической закономерностью (ЛЗ) класса $K_\lambda, \lambda=1,2,...,l,$ если

1. $\exists\mathbf{x}_t\in\widetilde{K}_\lambda: P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}_t)=1,$
2. $\exists\mathbf{x}_t\in\widetilde{K}_\mu, \mu=1,2,...,l, \mu\neq\lambda: P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}_t)=0,$

$\Phi(P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}))=\underset{\{P^{\Omega',\mathbf{c}^{1*},\mathbf{c}^{2*}}(\mathbf{x})\}}{extr}\Phi(P^{\Omega',\mathbf{c}^{1*},\mathbf{c}^{2*}}(\mathbf{x}))$, где Φ - критерий качества предиката.

Будем далее использовать стандартный критерий F качества класса $K_\lambda, \lambda=1,2,...,l$: $F(P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}))=\left|\{\mathbf{x}_i:\mathbf{x}_i\in\widetilde{K}_\lambda, P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}_i)=1, i=1,2,...,m\}\right|$ и понятия эквивалентных ЛЗ (принимающих равные значения на объектах обучающей выборки) и интервалов $N(P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}))$ ЛЗ $P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x})$. Пусть $\mathrm{X}(P^{\Omega,\mathbf{c}^1,\mathbf{c}^2})=\{\mathbf{x}_t\in K_\lambda: P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}_t)=1\}$. Будем называть $P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x})$ минимальной, если не существует ей эквивалентная ЛЗ $P^{\Omega_0,\mathbf{c}^3,\mathbf{c}^4}(\mathbf{x})$, для которой $N(P^{\Omega_0,\mathbf{c}^3,\mathbf{c}^4}(\mathbf{x}))\subset N(P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}))$. Если $P^{\Omega,\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x})$ является минимальной ЛЗ класса, то $\Omega=\{1,2,...,n\}, c_j^1=\min_t\{x_{tj}:\mathbf{x}_t\in\mathrm{X}(P^{\Omega,\mathbf{c}^1,\mathbf{c}^2})\},\ c_j^2=\max_t\{x_{tj}:\mathbf{x}_t\in\mathrm{X}(P^{\Omega,\mathbf{c}^1,\mathbf{c}^2})\}$.

Для минимальных ЛЗ введем понятия граничного объекта и множества граничных объектов. Поскольку далее модели распознавания будут основаны на нахождении и использовании минимальных ЛЗ, в обозначениях ЛЗ мы будем опускать символ Ω.

Определение. Объект $\mathbf{x}_t$ называется граничным для минимальной ЛЗ $P^{\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x})$, если $P^{\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}_t)=1$ и он является граничным для множества $N(P^{\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}))$. Множество граничных объектов обучения ЛЗ $P^{\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x})$ будем обозначать $G(P^{\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}))$. Множество неграничных (внутренних) объектов ЛЗ $P^{\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x})$ будем обозначать $L(P^{\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}))=\mathrm{X}(P^{\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}))\setminus G(P^{\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}))$.

Пусть $\mathrm{P}_j=\{P_{ji}^{\mathbf{c}^{1i},\mathbf{c}^{2i}}(\mathbf{x})\}-$ множество ЛЗ класса K_j.

Вычислим оценочную матрицу $\|V\|_{lxm}$ следующим образом: $V(i,j)=\sum_{r=1}^{|\mathrm{P}_j|} g_i^r(\mathbf{x}_j)$, где

$$g_i^r(\mathbf{x}_j)=\exp(-\sum_{k=1}^{n}(x_{jk}-\mu_{ik}^r)^2/2{\delta_{ik}^r}^2),\ \mu_{ik}^r=(\max_{x_j\in L(P_r)}x_{jk}+\min_{x_j\in L(P_r)}x_{jk})/2,$$

$$\delta_{ik}^r=(\max_{x_j\in L(P_r)}x_{jk}-\min_{x_j\in L(P_r)}x_{jk})/2,\ \text{если } L(P_r)\neq\varnothing, \text{ и } g_i^r(\mathbf{x}_j)=0 \text{ в противном случае.}$$

Будем относить объект обучения $\mathbf{x}_t$ к некоторому классу K_j по следующему правилу:

1. Если $\exists P_i^{\mathbf{c}^{1i},\mathbf{c}^{2i}}\in\mathrm{P}_j: P_i^{\mathbf{c}^{1i},\mathbf{c}^{2i}}(\mathbf{x}_t)=1,\mathbf{x}_t-$ внутренний объект для $N(P^{\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}))$, то $\widetilde{A}_t^y(\mathbf{x}_t)=j$;
2. Если $\forall P_i^{\mathbf{c}^{1i},\mathbf{c}^{2i}}\in\mathrm{P}_j: P_i^{\mathbf{c}^{1i},\mathbf{c}^{2i}}(\mathbf{x}_t)=1,\mathbf{x}_t\in G(P^{\mathbf{c}^1,\mathbf{c}^2}(\mathbf{x}))$, то $\widetilde{A}_t^y(\mathbf{x}_t)=\arg\max_{j=1,2,...,l}V(j,t).$

В случае, если максимум функции $V(j,t)$ достигается на нескольких значениях j, то берем из них произвольное.

Для решения задачи оптимизации (2) предлагается следующий алгоритм приближенного пересчета множества ЛЗ $P^j, j=1..l$, а также матрицы $\|V\|_{lxm}$ при рассмотрении соседней точки произвольного шага алгоритма оптимизации.

Зафиксируем некоторое значение вектора $\mathbf{y}$ и рассмотрим соседние для него точки по координате y_i.

$$\mathbf{y}^*=(y_1^*,y_2^*,...,y_{l-1}^*):\begin{cases}y_j^*=y_j, & j\neq i,\\ y_j^*\neq y_j, & j=i.\end{cases}$$

Без ограничения общности будем считать, что $y_i^*<y_i$. В таком случае некоторый объект $\mathbf{x}_t\in K_i$ перейдет из одного класса в другой, т.е. $K_j^*=K_j, j\neq i,i+1$, $K_i^*=K_i\setminus\{\mathbf{x}_t\}$, $K_{i+1}^*=K_{i+1}\cup\{\mathbf{x}_t\}$,

Обозначим через $\mathrm{P}_j^*, j=1,2,...,l,$ и $\|V\|_{lxm}$ набор ЛЗ и оценочную матрицу, соответствующие классам $K_1^*,K_2^*,...,K_l^*$. Многие ЛЗ и значения матрицы не изменяются: $\mathrm{P}_j^*=\mathrm{P}_j, j\neq i,i+1$, $V^*(j,t)=V(j,t), j\neq i,i+1,t=1,2,...,m.$

Обозначим множество объектов обучения, для которых $P_{ji}^{\mathbf{c}^{1i},\mathbf{c}^{2i}}(\mathbf{x}_t)=1$ через $\mathrm{X}(P_{ji})$. Для вычисления «приближений» $\mathrm{P}_j^*, j=1,2,...,l,$ и $\left\|V^*\right\|_{lxm}$ в режиме скользящего контроля предлагается следующий алгоритм.

Поочередно будем перебирать все ЛЗ P_{ji} множества $\mathrm{P}_j^*, j=1,2,...,l,$ и удалять объект $\mathbf{x}_t$ из каждой ЛЗ. Если $P_{ji}(\mathbf{x}_t)=0$, тогда $P_{ji}^*(\mathbf{x}_t)=P_{ji}(\mathbf{x}_t)$. Если $P_{ji}(\mathbf{x}_t)=1$, тогда рассмотрим 3 случая:

1. $\mathbf{x}_t$ является единственным объектом на одной из границ ЛЗ.

Тогда $\mathrm{X}(P_{ji}^*) = \mathrm{X}(P_{ji})\setminus\{\mathbf{x}_t\}$. Пересчитаем ЛЗ $P_{ji}^{*\mathbf{c}_j^1,\mathbf{c}_j^2}(\mathbf{x})=P_{ji}^{\hat{\mathbf{c}}_j^1,\hat{\mathbf{c}}_j^2}(\mathbf{x})$, где $\hat{c}_{j\alpha}^1=\min\limits_{\mathbf{x}_q\in\mathrm{X}(P_{ji}^*)} x_{q\alpha}$, $\hat{c}_{j\alpha}^2=\max\limits_{\mathbf{x}_q\in\mathrm{X}(P_{ji}^*)} x_{q\alpha}$, $\alpha=1,2,...,n.$

Пересчитываем $L(P_{ji}^*(\mathbf{x}))$ исходя из «новой» ЛЗ $P_{ji}^*(\mathbf{x})$, $V^*(i,t)=V(i,t)-g_j^r(\mathbf{x}_t)+g_j^{*r}(\mathbf{x}_t)$.

2. Если для некоторой ЛЗ P_{ji} имеет место $\mathbf{x}_t\in\mathrm{X}(P_{ji})$, то разбиваем $\mathrm{X}^*(P_{ji})$ на два множества по условиям $x_{q\alpha}>x_{t\alpha}$ и $x_{q\alpha}<x_{t\alpha}$, соответственно.
3. Если $\mathrm{X}(P_{ji})$ после удаления объекта $\mathbf{x}_t$ нельзя разбить на два, то рассматриваем последовательно все объекты $\mathrm{X}(P_{ji})$. Если рассматриваемый объект уже принадлежит какой-то другой ЛЗ класса, то множества ЛЗ не меняются. Если рассматриваемый объект (или несколько) не принадлежат никакому из других $\mathrm{X}(P_{ji}), i=1,2,...,\left|\mathrm{P}_j\right|$, то пытаемся расширить все $\mathrm{X}(P_{ji})$ за счет включения этого объекта. Если объект не может быть включить ни в одно из ранее имеющихся множеств $\mathrm{X}(P_{ji})$, то строим на объекте новую ЛЗ. Таким образом, вместо множеств ЛЗ $\mathrm{P}_j, j=1,2,...,l,$ вычисляем новые множества $\mathrm{P}_j^*, j=1,2,...,l$. Соответственно, вычисляем $V^*(i,t)$.

Таким образом, для алгоритма голосования по логическим закономерностям вычисление функционала в соседней точке общего алгоритма оптимизации также осуществляется эффективно с помощью «упрощенного» приближенного пересчета множеств ЛЗ.

Результаты работы модифицированного алгоритма вычисления оценок на практических данных

Результаты работы модифицированного алгоритма вычисления оценок сравнивались с алгоритмом линейной и квадратичной регрессии, регрессионного бинарного дерева принятия решений и двухслойной нейронной сети на реальных данных.

В качестве реальных данных рассматривалась задача “Relative CPU Performance Data” [Phillip Ein-Dor, Jacob Feldmesser, 1987] (прогнозирование производительности процессоров). Обучающая выборка состоит из 209 объектов, 8 признаков, из которых 2 признака в эксперименте не использовались (название производителя процессора и модель процессора). Таким образом, использовались 6 целочисленных признаков со следующими характеристиками:

Название признака	Минимальное значение	Максимальное значение	Среднее значение
Машинное время цикла	17	1500	203.8
Минимальная память	64	32000	2868.0
Максимальная память	64	64000	11796.1
КЭШ-память	0	256	25.2
Минимальное число каналов в процессоре	0	52	4.7
Максимальное число каналов в процессоре	0	176	18.2

Ниже на рисунках 1-2 показаны результаты работы алгоритмов на обучающей выборке. Модифицированный алгоритм вычисления оценок дает наилучший результат при числе классов l=64 (результат показан на рисунках 1 и 2). На рисунках 3-4 показаны результаты работы алгоритмов в режиме скользящего контроля leave-one-out (LOO).

Обозначения точек и графиков

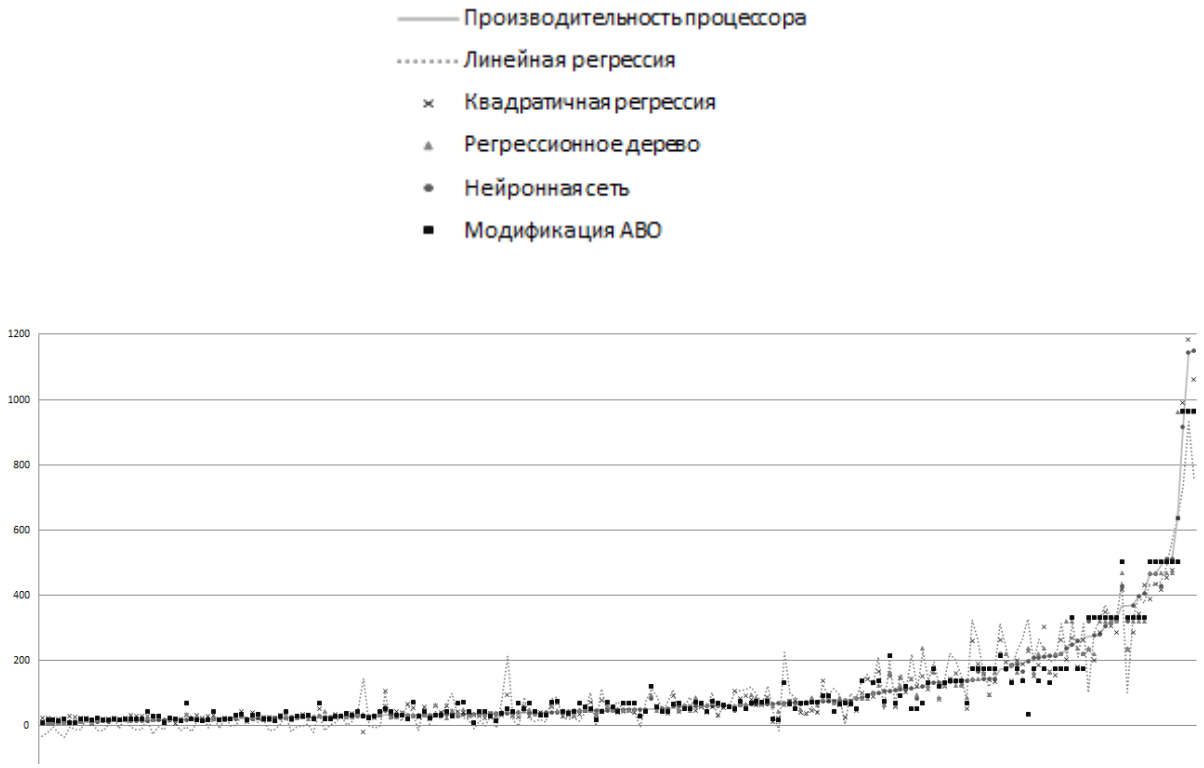


Рис. 1. Результаты прогнозирования на обучающей выборке

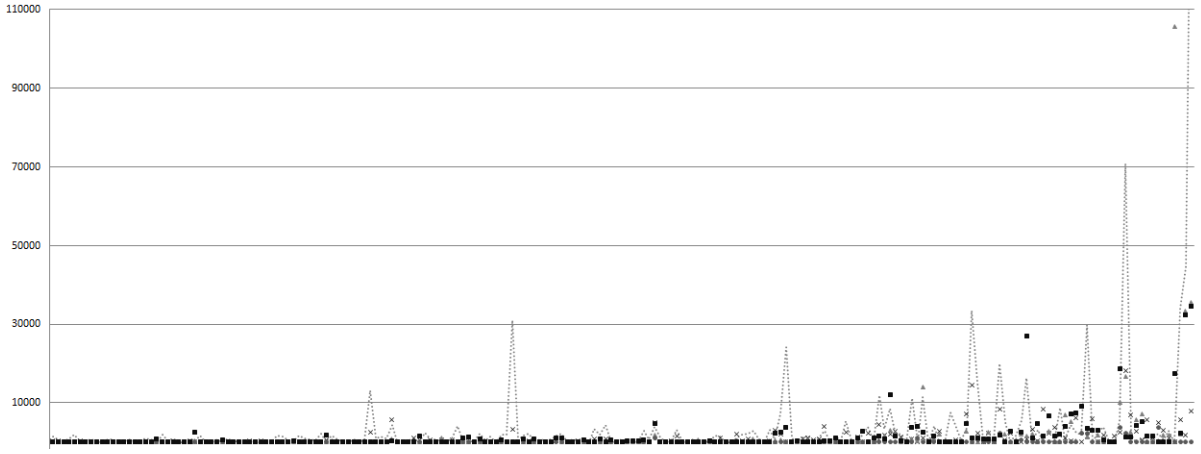


Рис. 2. Квадратичная ошибка на обучающей выборке

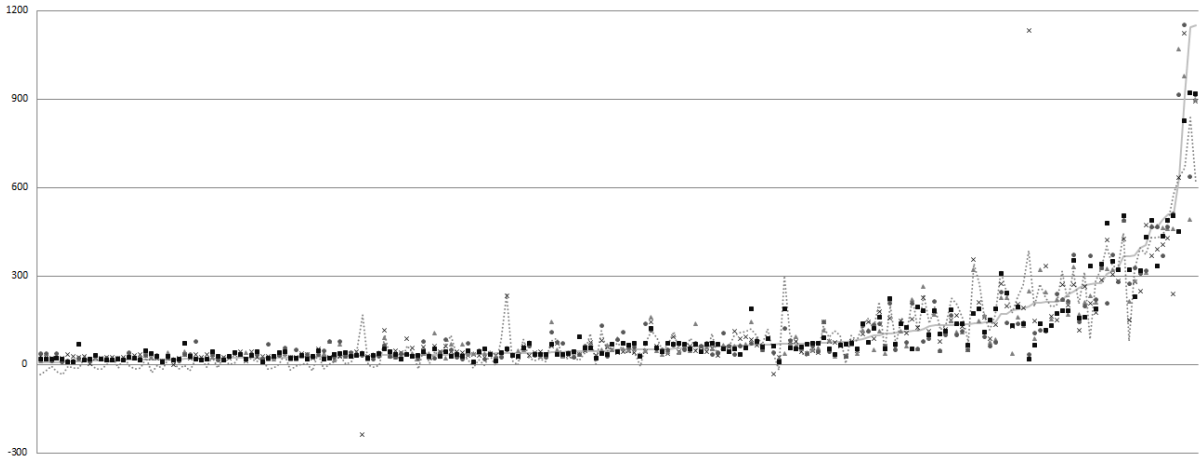


Рис. 3. Результаты прогнозирования в режиме скользящего контроля

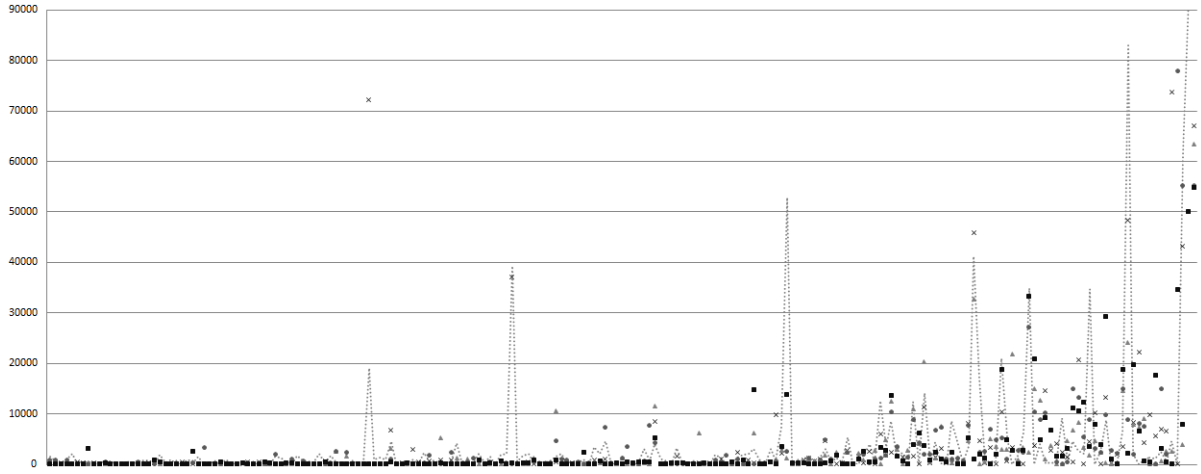


Рис. 4. Квадратичная ошибка на обучающей выборке в режиме скользящего контроля

Суммарные значения квадратичной ошибки алгоритмов по всем объектам

	Ошибка Модификации АВО	Ошибка линейной регрессии	Ошибка квадратичной регрессии	Ошибка регрессионного дерева	Ошибка нейронной сети
На обучающей выборке	295 822,56	762 326,89	220 317,75	304 987,35	20 827,23
В режиме скользящего ронтроля LOO	534 275,125	1 081 745,973	3 092 982,792	1 047 395,129	819 349,2052

При скользящем контроле модифицированный алгоритм вычисления оценок показал меньшее суммарное значение квадратичной ошибки, чем алгоритмы линейной и квадратичной регрессии, регрессионного дерева и двухслойной нейронной сети.

Заключение

Представляется важным отметить следующие детали настоящего подхода.

1. Алгоритмы восстановления регрессии рассматривались для случая числовых признаков, $x_i \in R, i = 1,2,...,n$. Легко видеть, что при построении кусочно-постоянных регрессий по обучающим выборкам признаки могут быть разнотипными: числовыми, бинарными, порядковыми. Предложенная модель восстановления регрессии основана на использовании модификации модели вычисления оценок, которая не требует наличия метрики в признаковом пространстве.
2. В начале работы отмечены некоторые общие случаи, когда классические методы восстановления регрессий плохо применимы. Здесь следует добавить случаи, когда значения зависимой величины представлены весьма неравномерно, или она является фактически k -значной величиной с большим значением k . Для задач распознавания наиболее предпочтителен случай с минимальным числом классов 2, особенно при фиксированной длине выборки. Таким образом, ожидается, что модель построения кусочно-постоянных регрессий будет полезна при решении многочисленных «плохих» задач, неудобных как для стандартных регрессионных подходов, так и для задач распознавания. В данных случаях «плохая» задача регрессии сводится к задаче распознавания с оптимальным выбором k .

В модели построения кусочно-постоянных регрессий мы полагали, что $f(\mathbf{x}) = (a_j + a_{j-1})/2$, $j = 1,2,...,l$, $a_0 = a$, $a_l = b$. Здесь конечно возможны и более точные способы непараметрического оценивания функций. Представляется перспективным их вычисление, согласованное с вычислением функций близости (8), или использование модели распознавания, основанной на голосовании по системам логических закономерностей классов [Рязанов, 2007].

3. Построение кусочно-квадратичных, кусочно-полиномиальных, и т.п. функций может осуществляться аналогично построению кусочно-линейных функций.

Acknowledgements

Настоящая работа выполнена при поддержке Программы Президиума РАН №14 , Программы №2 Отделения математических наук РАН, проектов РФФИ № 11-01-00585-а , № 12-01-90012-бел-а , № 12-01-00912-а , N 13-01-90616 -арм-а.

Библиография

[Bellman, 1957] R. Bellman, "Dynamic programming", Princeton Univ. Press (1957)

[Collobert, Bengio, 2001] R. Collobert, S. Bengio. SVMTorch: Support Vector Machines for Large-Scale Regression Problems. Journal of Machine Learning Research, 1:143-160, 2001.

[Jing-Rung Yu, Gwo-Hshiung Tzeng, Han-Lin Li 2001] Jing-Rung Yu, Gwo-Hshiung Tzeng, Han-Lin Li: General fuzzy piecewise regression analysis with automatic change-point detection. Fuzzy Sets and Systems 119(2): 247-257 (2001)

[Дрейпер, Смит, 2007] Дрейпер Н., Смит Г. Прикладной регрессионный анализ. М.: Издательский дом Вильямс, 2007.

[Журавлев, Никифоров, 1971] Журавлев Ю.И., Никифоров В.В. Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. 1971. №3. С. 1-11.

[Ковшов, Моисеев, Рязанов, 2008] Ковшов Н.В., Моисеев В.Л., Рязанов В.В. Алгоритмы поиска логических закономерностей в задачах распознавания // Журнал вычислительной математики и математической физики, Т.48, 2008, N 2, стр. 329-344.

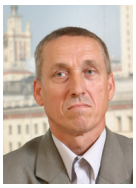
[Михалевич, 1965] Михалевич В.С. Последовательные алгоритмы оптимизации и их применение// Кибернетика. 1965, №1. С.45-46.

[Рязанов, 2007] Рязанов В.В. Логические закономерности в задачах распознавания (параметрический подход) // Журнал вычислительной математики и математической физики, Т.47, №10, 2007, с.1793-1808

[Рязанов, Тишин, Щичко, 2009] Рязанов В.В., Тишин К.В., Щичко А.С. Восстановление зависимостей по прецедентам на основе применения методов распознавания и динамического программирования// Математические методы распознавания образов: 14-я Всероссийская конференция. Владимирская обл., г.Суздаль, 21-26 сентября 2009 г.: Сборник докладов.-М.: МАКС Пресс, 2009. Стр. 168-171

[Хардле, 1993] Хардле В. Прикладная непараметрическая регрессия. М., Мир, 1993.

Информация об авторах



Vladimir Ryazanov *– Heard of Department; Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Russia, 119991 Moscow, Vavilov's street, 40;*

e-mail: rvv@ccas.ru

Major Fields of Scientific Research: Pattern recognition, Data mining, Artificial Intelligence

Anton Shchichko *– Postgraduate student; Lomonosov Moscow State University; Faculty of Computational Mathematics and Cybernetics; Department of Mathematical Forecasting Methods, Russia, 119991, Moscow, Kravchenko, 7;*

e-mail: anton.schichko@gmail.com

Major Fields of Scientific Research: Pattern recognition, Data mining, Artificial Intelligence

ЛОГИКО-ЛИНГВИСТИЧЕСКАЯ МОДЕЛЬ ИЗВЛЕЧЕНИЯ ФАКТОВ ИЗ СЛАБОСТРУКТУРИРОВАННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

Нина Хайрова, Наталья Шаронова

***Аннотация**: Одним из перспективных направлений информационного поиска является фактографический поиск и разработка фактографических баз данных. Существующие сегодня модели и алгоритмы фактографического поиска в своем большинстве направлены на излечение фактов из хорошо формализованной информации, в том числе из хорошо формализованной текстовой информации. В работе предлагается модель извлечения фактографической информации из динамически меняющихся слабоформализованных текстовых потоков, не ограниченных определенными предметными областями. Для извлечения некоторого факта используется шаблон «агент-предикат-значение», отображающий отношения, формально выражаемые семантическими падежами партиципантов предложения. В предлагаемой логико-лингвистической модели семантические роли именных групп определяются отношением четко выделенных множеств морфологических, синтаксических и семантически категорий, описываемым с помощью базового аппарата алгебры конечных предикатов. В работе рассмотрена реализация данной модели для извлечения фактографической информации о дате, месте рождения и роде деятельности персоналии из русскоязычных слабоформалзованных текстов. Экспериментальная проверка программной имплементации модели показала правильность выделения факта примерно в 94,3% случаев.*

***Ключевые слова**: фактографический поиск, слабоформализованная текстовая информация, лингвистический процессор, алгебра конечных предикатов.*

***ACM Classification Keywords**: H.3.3 .Information Search and Retrieval*

Введение

Одним из главных приложений Natural Language Processing (NLP) является информационный поиск, классифицируемый по результатам выдачи на документальный поиск – процесс поиска документа в массивах первичных или вторичных документов, и фактографический поиск – процесс поиска фактов, отвечающих информационному запросу. Факт представляет собой знание в форме утверждения, достоверность которого строго установлена [Барахнин, 1980]. На практике, в сфере информационных технологий, фактографическую информацию обычно трактуют несколько иначе — как конкретные сведения или данные независимо от того, являются ли они фактическими или прогнозируемыми. Главное, что эти сведения сообщают о какой-то предметной области, а не о документах, посвященные этой области.

В свою очередь фактографическую информацию можно разделить на хорошо структурированную и плохо структурированную. К хорошо структурированным сведениям (так называемая параметрическая информация) относятся, прежде всего, сведения количественного характера, а так же качественные сведения, имеющие хорошо регламентированную форму. К плохо структурированной фактографической

информации относятся сведения, представленные различными нерегламентированными словесными конструкциями, представленными на естественном языке [Baeza-Yates, 1999].

Алгоритмы фактографического анализа зависят, в свою очередь, от степени структурированности конкретного документа [Ландэ, 2009]. По степени структурированности данные документа можно разделить, подобно общей классификации степени формализации информации, на табличные данные, отображенные в виде фактов; массивы однородных слабоструктурированных текстовых документов, обычно описывающие конкретную предметную область, и документы произвольного слабоструктурированного типа.

Актуальность исследования

Большое влияние на разработку методов и моделей извлечения фактографической информации оказала серия конференций Message Understanding Conferences, MUC, проходившая с 1987 г. по 1997 г. при поддержке Американского агентства DARPA (Defense Advanced Research Projects Agency), и способствующая упорядочиванию информации по системам фактографического поиска. Но только в последние несколько лет стали появляться реальные системы, включающие элементы такого рода поиска. Например, практически единственная на территории СНГ, находящаяся в тестовом режиме, российская поисковая система nigma.ru. Главным ресурсом извлечения которой, являются слабоструктурированная wikipedia.org, как самая большая база текстов сходной структуры. Системы ask.com и answers.com так же являются системами фактографического поиска, выполняющие поиск в документах, а в качестве ответа на более общие вопросы, использующие ссылки на ресурс wikipedia.org.

Анализ показывает, что основные усилия разработчиков фактографических информационно-поисковых систем (ФИПС) направлены на хорошо структурированные факты, извлечение которых, с одной стороны, легче автоматизировать, а с другой, к этому типу относится почти вся производственно-экономическая информация, циркулирующая в сфере материального производства и управления. Таким образом, для извлечения фактов, представленных в табличных данных и в массивах однородной структурированной текстовой информации существуют достаточно надежные алгоритмы. Задача же извлечения фактов из произвольных текстов естественного языка до сих пор не имеет сколько-нибудь общего решения [Baeza-Yates, 1999], [Ландэ, 2009].

Проблема создания ФИПС, работающих со слабоформализованной текстовой информаций, еще далека от полного решения. Один из существующих подходов извлечения фактов из подобных текстов заключается в использовании онтологий или тезаурусов предметной области, которые и определяют, что является фактом, в рамках данной онтологии. Но такого рода подход, опять же, ограничивает анализируемые полнотекстовые документы узкой предметной областью онтологии.

Разработка методов и алгоритмов извлечения фактов из динамически меняющихся слабоструктурированных текстовых потоков, не ограниченных определенными предметными областями, требует точного моделирования когнитивной деятельности человека, по пониманию и идентификации фактов, а также наличия мощных средств как синтаксического, так и семантического анализа текстов, учитывающих семантическую эквивалентность и мультиязычность.

Общая постановка задачи

Опираясь на определение факта, можно дать определение минимальной смысловой единицы фактографического поиска, представляющую собой триаду: агент-предикат-значение (табл.1). То есть запись фактографической информации должна включать указатель на агента фактографического поиска,

на атрибут или предикат этого объекта и давать конкретное значение этого атрибута. Такое определение позволяет извлекать понятия из слабоструктурированных текстовых источников информации и представлять отношения между ними в структурированном виде. Получаемая структура представляет собой факты, как в виде достаточно простых понятий: ключевых слов, персоналий, организаций, географических названий, так и в более сложном виде, например, имя персоналии с ее должностью и родом деятельности.

Таблица 1. Примеры формального представления фактической информации

Агент	*Атрибут*	*Значение*
пациент Иванов	резус-фактор	положительный
системный блок	вес	5 кг
Афродита	место рождения	Кипр

В общем случае выделение фактов из слабоструктурированной текстовой информации включает следующие этапы:

1) Entity Extraction – извлечение слов или словосочетаний, важных для описания смысла текста (списки терминов предметной области, персоналий, организаций, географических названий и т.д);
2) Feature Association Extraction – исследование связей между извлеченными понятиями;
3) Event and Fact Extraction – извлечение сущностей, распознавание фактов и действий.

Для реализации первого этапа выделения понятий используется стандартный лингвистический процессор [Хайрова, 2010], включающий графемную, морфологическую, синтаксическую и контекстную этапы обработки, с добавлением специализированных методов и алгоритмов обработки документов. Так как очень часто в задачах по извлечению фактографической информации нужно найти в тексте упоминания лиц, компаний, правительственных организаций и местоположений, и другие подобные типы сущностей, то для их выделения используются специальные формализмы графемного анализа. На этапе морфологического анализа используется декларативный и алгоритмический методы. Каждый неправильный глагол английского языка представлен в базе данных во всех его формах, то есть глагол *write* иметь формы *write-writes-wrote-written-writing*, формы правильных глаголов определяются алгоритмически. Русское словоизменение определяется алгоритмически, с использованием словарей окончаний.

А так как для построения триады фактографической информации необходимо выделить сущности, представленные в текстах под разными именами, то особое значение приобретает этап разрешения кореферентности (coreference resolution) синтаксического анализа, для определения синонимов, интересуемых сущностей. На этом этапе такие местоимения как «он», «она», «они», «ним» и т.д. должны быть ассоциированы со своими антецедентами, соотнесены их с именуемой сущностью данной предметной области.

Центральной задачей получения фактографической информации является второй этап обработки, представляющий извлечение информации об отношениях между сущностями. При этом, для извлечения некоторого факта необходимо определить некий шаблон, отображающий семантические (или понятийные) связи в предложении. Для задания таких смысловых отношений предлагается использовать грамматику семантических падежей. Для чего необходимо разработать строгую модель, связывающую информацию, содержащуюся в определении семантических ролей с элементами поверхностной структуры предложений естественного языка. Такой подход рассматривается в рамках падежной грамматики и основывается на понятии глубинных падежей, введенных Ч. Филлмором, выделившим

пропзицию, или основной смысл предложения, как предикат, выражаемый в поверхностной структуре чаще глаголом, связанным с помощью определенных глубинных падежей с участниками данной ситуации, или партиципантами [Филлмор, 1981].

Для того, чтобы получить возможность использования глубинных падежей в задачах автоматической обработки смысла, необходимо формально определить глубинный семантический падеж через поверхностную структуру предложения данного языка. Семантические падежи в различных естественных языках имеют разные формы формального выражения. Например, в русском и украинском языках, семантическая информация партиципантов кодируется, в основном, грамматическими поверхностными падежами, тогда как в английском — она передается либо сочетанием с предлогом, либо, порядком слов в предложении. Но сегодня, в связи с отсутствием четко сформулированных критериев выделения семантических ролей, отсутствуют формальные модели, содержащие полные наборы семантических ролей, с достаточной степенью точности выражаемых элементами поверхностной структуры предложения, что приводит к низкому уровню использования данного подхода при автоматической обработке текстов.

Описание математической модели

Введем на универсуме U, включающем все возможные понятия и объекты анализа сложной языковой системы [Хайрова, 2012], множество грамматико-семантических характеристик лексем предложения $M = \{m_1, \dots, m_n\}$, где n — количество характеристик системы. Используя формальный аппарат алгебры конечных предикатов [Бондаренко, 2007] , формально определим отношения между морфологическими и семантическими характеристиками существительного, формально выражающими семантические падежи партиципантов предложения русского языка. Для этого представим отношения между характеристиками в виде $m_i * m_j$, *где* $m_i, m_j \in M$, а знак $*$ — обозначает, что данные характеристики соответствуют существительному, относящемуся к определенному семантическому падежу. На множестве M введем систему предикатов S так, чтобы любой предикат $P(q_m) \in S$, обращался в *1* на множестве лексем с грамматико-семантической информацией, соответствующей определяемому семантическому падежу и был равен *0* в противном случае, сопоставляя множество предикатов S множеству семантико-грамматических характеристик приписанных лексеме.

Количество и состав семантических ролей, и предметных переменных, выделяемых при описании языка, могут существенно различаться в зависимости от задач описания, языка и его степени детализации [Филлмор, 1981]. Для формального определения семантических падежей русского языка выделим достаточно четко сформированное множество семантико-грамматических признаков, с помощью несократимого набора трех переменных: X-признак одушевленности (со значениями $x^о$ — предметная переменная, характеризующая семантический признак живого, $x^н$ — предметная переменная характеризующая семантический признак неживого;); Y –элемент семантического значения существительного ($y^м$ — механизм, $y^с$ — имя собственное, $y^и$ — инструмент, $y^ч$ — часть тела, $y^т$ — плоскость/точка, $y^о$ — объемное пространство, $y^в$ — определенное время, $y^п$ — период, $y^ц$ — пункт назначения); Z — грамматический падеж существительного ($z^и$, $z^р$, $z^д$, $z^в$, $z^т$, $z^п$— предметные переменные, описывающие свойства существительных обладать тем или иным грамматическим падежом). Область изменения введенных переменных формально задается следующим образом:

$$x^о \vee x^н = 1,$$
$$z^и \vee z^р \vee z^д \vee z^в \vee z^т \vee z^п = 1, \quad (1)$$
$$y^м \vee y^с \vee y^и \vee y^ч \vee y^т \vee y^о \vee y^в \vee y^п \vee y^ц = 1,$$

Семантический падеж существительного предложения определяются через предикат P, связывающий элементы семантического значения существительного *x* и *y* с его грамматическими значениями *z*:

$$P(x, y, z) \rightarrow P(x) \bullet P(y) \bullet P(z), \quad (2)$$

где • — операция конъюнкции. Так как возможность согласования морфосемантической информации не зависит от того, к какой конкретно словоформе она относится, на декартовом квадрате множества $S * S$ можно задать предикат $\gamma\ (x_n, y_n, z_n)$, принимающий значение *1*, если морфосемантическая информация словоформы n формирует некоторый семантический падеж лексемы, и значение *0* в противном случае. Таким образом, отношения морфосемантических признаков существительных предложения, выражающих семантические падежи, требуемые валентностью глагола, можно задать формулой:

$$P(x_n) * P(y_n) * P(z_n) = \gamma_k (x_n, y_n, z_n) \bullet P(x_n) \bullet P(y_n) \bullet P(z_n), \quad (3)$$

Практически никогда подмножество согласующейся морфосемантической информации, выражающей семантические падежи, не совпадает с декартовым произведением на множестве морфологических и семантических признаков. Те морфосемантические признаки, которые в своем согласовании не формируют семантический падеж существительного должны исключаться из формулы (2) множителями $\gamma_k (x_n, y_n, z_n)$, $k \in [1;m]$, где *m* — количество, принятых к рассмотрению в системе семантических падежей.

В соответствии с формулой (3) семантический падеж *агенс* определяется множеством возможных связей морфосемантической информации существительного:

$$P_1 (x_n, y_n, z_n) = (x_n^{o}\, z_n^{и} \vee z^{и}\, x_n^{н}\, y_n^{м} \vee z^{и}\, x_n^{o}\, y_n^{c})\ (y_n^{м} \vee y_n^{c} \vee y_n^{и} \vee y_n^{ч} \vee y_n^{т} \vee y_n^{o} \vee y_n^{в} \vee y_n^{п} \vee y_n^{ц})(x_n^{o} \vee x_n^{н})\ (z^{и}_n \vee z_n^{р} \vee z_n^{д} \vee z_n^{в} \vee z_n^{т} \vee z_n^{п}) \quad (4)$$

Множество возможных связей грамматической и семантической информации существительного семантического падежа *со-агенс* задается предикатом $P_2(x_n, y_n, z_n)$:

$$P_2 (x_n, y_n, z_n) = (z^{т}\, x_n^{o}\, y_n^{c})\ (y_n^{м} \vee y_n^{c} \vee y_n^{и} \vee y_n^{ч} \vee y_n^{т} \vee y_n^{o} \vee y_n^{в} \vee y_n^{п} \vee y_n^{ц})\ (x_n^{o} \vee x_n^{н})\ (z^{и}_n \vee z_n^{р} \vee z_n^{д} \vee z_n^{в} \vee z_n^{т} \vee z_n^{п}) \quad (5)$$

Множество возможных связей грамматической и семантической информации существительного семантического падежа *темпоралис* задается предикатом $P_3(x_n, y_n, z_n)$:

$$P_3 (x_n, y_n, z_n) = (z^{в}\, x_n^{н}\, y_n^{в} \vee z^{п}\, x_n^{н}\, y_n^{п})\ (y_n^{м} \vee y_n^{c} \vee y_n^{и} \vee y_n^{ч} \vee y_n^{т} \vee y_n^{o} \vee y_n^{в} \vee y_n^{п} \vee y_n^{ц})\ (x_n^{o} \vee x_n^{н})\ (z^{и}_n \vee z_n^{р} \vee z_n^{д} \vee z_n^{в} \vee z_n^{т} \vee z_n^{п}) \quad (6)$$

Множество возможных связей грамматической и семантической информации существительного семантического падежа *локатив* задается предикатом $P_4(x_n, y_n, z_n)$:

$$P_4 (x_n, y_n, z_n) = (z^{п}\, x_n^{н}\, y_n^{т} \vee z^{п}\, x_n^{н} y_n^{м} \vee z^{п}\, x_n^{н} y_n^{ц} \vee z^{п}\, x_n^{н} y_n^{o})\ (y_n^{м} \vee y_n^{c} \vee y_n^{и} \vee y_n^{ч} \vee y_n^{т} \vee y_n^{o} \vee y_n^{в} \vee y_n^{п} \vee y_n^{ц})\ (x_n^{o} \vee x_n^{н})\ (z^{и}_n \vee z_n^{р} \vee z_n^{д} \vee z_n^{в} \vee z_n^{т} \vee z_n^{п}) \quad (7)$$

Для английского языка в дополнение к семантическим признакам вместо морфологических категорий выбираются синтаксические признаки употребления предлогов, т.к. именно предлоги после глаголов выражают валентность глагола английского языка. Признаки могут быть как уникальными для определенных глаголов, так и общими, как, например, признак направления движения может быть выражен глаголами go, run, drive, ride, transport, ship и т.д. Признаки могут быть описаны, так же как и для объектов с помощью семантических и грамматических характеристик. Например, для семантического падежа *Локатива*, если именная форма имеет признак плоскости / точки, употребляется английский

предлог *on*, а если объемного пространства, то предлог *in*. Падеж *темпоралис* определяется в английском языке предлогами *in* и *on*. Предлог *in* выражает год, в котором произошло событие, предлог *on* выражает точную дату, например, *Hewas born on February, 7 1902*. Падеж *локатива* определяется предлогом *in*, при выражении места, где произошло событие. Падеж *фактитив*, не требует предлога после глагола, что и является главным требование (наличие после определенного слова предлога показывает, это глагол не требует падежа *фактитив*).

Структурное описание модели

Для извлечения из неструктурированной текстовой знаний или фактов необходимо определить семантические падежи, определяющие данные факты, глаголы языка, выражающие предикат в триаде факта «агент – предикат – значение» и предикаты морфосемантических (или семантико-синтаксических) признаков существительных предложения разработанной модели (3), выражающие семантические падежи объекта триады факта.

Для определения фактов о дате, месте рождения и виде деятельности персоналии в виде: «агент-предикат-значение» используем, кроме семантического падежа *агент*, три семантических падежа, выражающие информацию, соответствующую требованиям: *темпоралис* - временная характеристика события, позволяющая определить дату рождения человека; *локатив* - падеж, характеризующий местонахождение, положение или состояние объекта, позволяя определить место рождения человека; и *фактитив* - падеж, определяющий сферы и продукты деятельности человека. Нами был выбран именно *фактитив*, а не *объектив*, так как именно *фактитив* дает значение результата действия.

В процессе реализации модели был определен набор глаголов, требующих определенного семантического падежа партиципантов предложения. Для этого с использованием толковых и переводных словарей были проанализированы около 130 биографий ученых и деятелей и выделены наиболее распространенные в данном типе слабострурированной текстовой информации глаголы английского и русского языка, выражающие предикат требуемого действия (табл. 2). Количество и список глаголов может меняться в зависимости от цели и объектов (фактов) поиска.

Таблица 2. Семантические падежи и наиболее распространенные предикаты, соответствующие фактам даты, места рождения и деятельности некоторой персоналии.

Падеж	Английский предикат	Русский предикат	Пример
Темпоралис	born, came up, saw the light, came into the world	родиться	He was born on February, 7 1902.
Локатив	born, came up, saw the light, came into the world	родиться	She came into the world in Frankfurt.
Фактитив	write, produce, paint, draw, pencil, design, project, invent, discover, engineer, compose, create, construct, publish, establish, investigate, research, explore, contribute, make, describe, work и др.	писать, изготовлять, рисовать, проектировать, открывать, изобретать, разрабатывать, создавать строить, публиковать, учреждать, исследовать, вносить вклад, сделать, описывать, работать	There he wrote his first poem. In 1843 he completed the drama NazarStodolya.

Для определения соответствия информации определенному факту были заданы требуемые семантические падежи для каждого глагола. Например, русский глагол «*написать»,* аналогично английскому глаголу «*write»* формирует семантические падежи партиципантов представленные в табл. 3.

Таблица 3. Формальное определение требуемых семантических падежей глагола «*написать*» («*write*»)

towritetosmb	адресат
towritesmth	объектив (работа, рукопись, издание)
towritewithsmth	инструменталис (инструмент письма)
tobewrittenbysmb	агент (автор)
whenwritein	темпоралис (время написания)
ftowritein	локатив (место написания)

Осуществление детального анализа текстов биографий позволило определить дополнительные условия выражения семантических падежей, определяющий дату, место и род занятий персоналии:

- семантический падеж *локатив* выражается именем собственным (обычно графически выражаемым большой буквой), так как нас интересует населенный пункт, а не местоположение, как например *in mansion*;
- семантический падеж *фактитив* допускает несколько вариантов выражения: значащее слово написано с большой буквы, слово или словосочетание взято в скобки, или оно является существительным.

Программная имплементация модели

Программная имплементация модели представляет собой веб-приложение, анализирующее текст или список анализируемых текстовых файлов. При нажатии кнопки *checkname* (рис. 1) программа выделяет первые 2 слова с большой буквы в первом предложении и выводит их на экран, как возможные, имя и фамилия лица, о котором идет речь. Так как правила написания биографий, представляющих слабоформализованные тексты на естественном языке, практически, одинаковы, было определено, что первые 2 слова с большой буквы, с вероятность 99,5, определяют имя и фамилию персоналии, о которой идет речь в данной библиографии. В случае, если выделенные данные не верны, пользователь может сам вписать имя, это понадобится в разборе биографии *ПаблоПикасо*, имеющего полное имя *Pablo Diego José Francis code Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Ruiz y Picasso*. Результат работы программы представлены в виде диалогового окна,

Извлеченная системой фактографическая информация представляется пользователю форме диалогового окна (см. рис.1). Программа отображает извлеченную фактографическую информацию в виде факта и первичных предложений, из которых данный факт был извлечен. В поле *birthday* находится информация даты рождения - *7 February 1812* и предложения, из которых была извлечена данная информация. Аналогичным образом в полях *origin* и *activity* отображается фактографическая информация о месте рождения и роде деятельности. Если в полях *birthday* и *origin* существует лишь один верный ответ, то поле *activity* может содержать несколько истинных ответов. Информация о деятельности записывается последовательно, каждый факт с нового абзаца. Факты деятельности располагаются в порядке значимости, определенной системой.

Экспериментальная проверка, проведенная на 47 полнотекстовых библиографических текстах электронного фонда ХГНБ, показали правильность определения даты рождения – 97,9% случаев, правильность определения места рождения – 95,7% , деятельность персоналии – 89,4%.

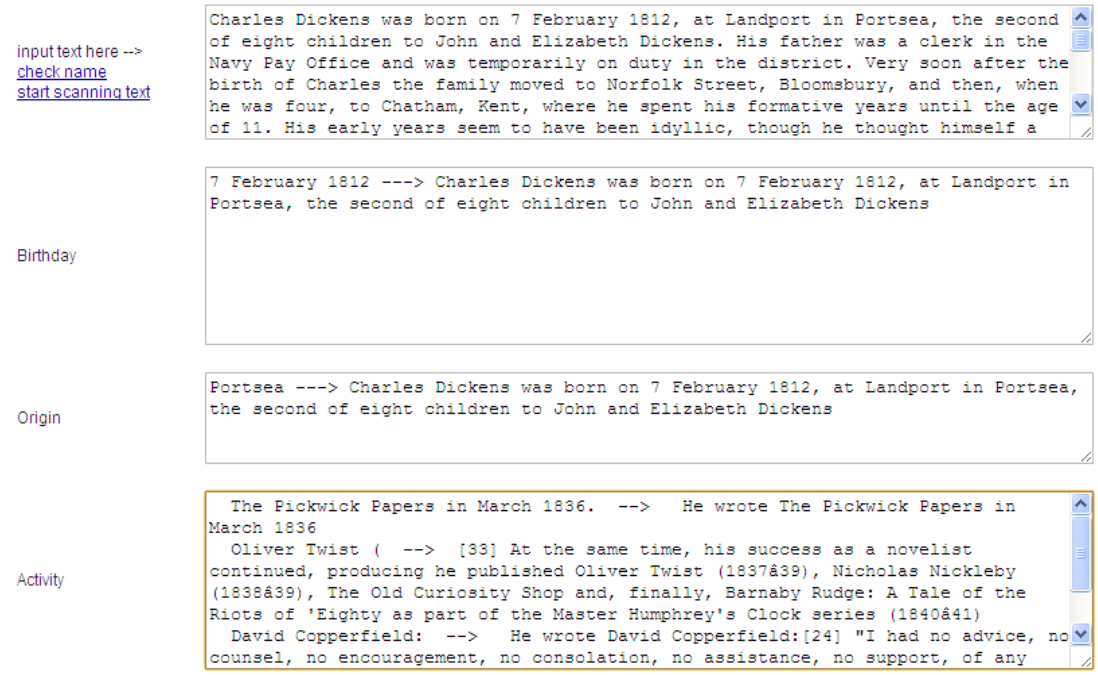


Рис. 1. Окно результата извлеченной фактографической информации рассматриваемой модели.

Выводы

Результатом данного исследования является разработка логико-лингвистической модели извлечения фактов из слабоструктурированных текстов, не ограниченных определенными предметными областями. Для извлечения факта использован шаблон «агент-предикат-значение», отображающий семантические (смысловые) отношения в предложении. Для задания таких смысловых отношений используется грамматика семантических падежей. В предлагаемой логико-лингвистической модели семантические падежи именных групп определяются предикатами, выражающими отношения четко выделенных множеств морфологических, синтаксических и семантически категорий. Множество возможных связей грамматической и семантической информации партиципантов предложения описано средствами аппарата алгебры конечных предикатов. В работе рассмотрена реализация модели для извлечения фактографической информации о дате, месте рождения и роде деятельности персоналии из русскоязычных слабоформализованных текстов. Экспериментальная проверка программной имплементации модели показала правильность выделения факта примерно в 94,3% случаев.

Литература:

[Baeza-Yates, 1999] Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval. — Addison-Wesley, 1999. — 340 p.

[Барахнин, 1980] Барахнин В.Б., Федотов А.М.Проблемы разработки технологии фактографического поиска – М.: Институт вычислительных технологий СО РАН, 1980. – 150 с.

[Бондаренко, 2007] Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П. Теория интеллекта. Харьков. Изд-во СМИТ. 2007. 576 с.

[Ландэ, 2009] Ландэ Д.В., Снарский А.А., Безсуднов И.В. Интернетика: Навигация в сложных сетях: модели и алгоритмы - М.: Либроком (Editorial URSS), 2009. - 264 с.

[Ушаков, 1939] Толковый словарь русского языка: В 4 т. / Под ред. Д. Н. Ушакова. Т. 1. М., 1935; Т. 2. М., 1938; Т. 3. М., 1939; Т. 4, М., 1940.

[Филлмор, 1981] Филлмор Ч. Дело о падеже открывается вновь // Новое в зарубежной лингвистике. – М.: Изд. иностр.лит., 1981, вып. 10. – С. 496-530

[Хайрова, 2010] Хайрова Н. Ф.. Тарловский В. А. Использование семантико-ориентированного лингвистического процессора для добывания новых знаний из потока документов корпоративной информационной системы / Вісник Національного технічного університету «ХПІ». Збірник наукових праць. Мематичний випуск «Системнй аналіз, управління та інформаційні технології». — Х.: НТУ «ХПІ». — 2010. — № 67. — С. 132-138.

[Хайрова, 2012] Хайрова Н. Ф. Використання логіко-алгебраїчної моделі семантичних відмінків для семантичного аналізу речення/ Н. Ф. Хайрова // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2012.- Вип. № 38. – С. 239 – 245.

Сведения об авторах

Хайрова Нина *– доцент кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», ул. Фрунзе, 21, Харьков, 61002, Украина; e-mail: nina_khajrova@yahoo.com*

Научные интересы: искусственный интеллект, обработка знаний, автоматическая обработка текстов

Шаронова Наталья *– профессор, заведующий кафедрой интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», ул. Фрунзе, 21, Харьков, 61002, Украина; e-mail:, nvsharonova@mail.ru*

Научные интересы: искусственный интеллект, математическое моделирование, автоматизированные библиотечные системы

ВЫДЕЛЕНИЕ ТЕКСТА НА СЛОЖНОМ ЦВЕТНОМ ФОНЕ

Роман Телятников, Иван Шумский, Ариф Мамедов, Анатолий Протосавицкий, Екатерина Матусевич, Екатерина Степанькова

***Аннотация**: В работе исследовано решение прикладной задачи выделения текста на сложном цветном фоне. Разработана эффективная процедура фильтрации фона от текста путем свертки цветного изображения в полутоновое. В основу процедуры фильтрации положена теория классификации. С целью повышения эффективности фильтрации предложен способ компенсации цветовых искажений, вызванных аппаратурой сканирования. Для практического применения процедуры выработаны конкретные рекомендации по выбору цветового пространства и ядра функции расстояния.*

***Ключевые слова**: фильтрация, цветовая модель, гистограмма, функция расстояния.*

***ACM Classification Keywords:** CCS - Computing methodologies - Computer graphics - Image manipulation - Image processing; CCS - Applied computing - Document management and text processing - Document capture - Optical character recognition.*

Введение

Существует огромный перечень прикладных задач, в которых на изображении требуется качественно выделить объект на сложном фоне. Если речь идет об обработке одного изображения, то для этой цели вполне пригодны профессиональные графические редакторы (Adobe Photoshop, Corel Photo-Paint, GIMP, и др. [Wikipedia, 2013]). В этих редакторах реализованы инструменты типа «Свободное выделение (Лассо)», «Умные ножницы» и «Выделение переднего плана», позволяющие вручную выполнить фильтрацию.

В случае необходимости обработки множества изображений, например, при обработке кадров видеоряда или при оцифровке сканов однотипных документов, требуется автоматизация процесса фильтрации. При этом допускается настройка параметров фильтра на одном изображении, но при обработке остальных изображений параметры должны автоматически подстраиваться под изменяющиеся цветовые характеристики объекта и фона.

Задача формулируется следующим образом: необходимо разработать процедуру фильтрации, увеличивающую расстояние между цветом объекта и цветом фона в RGB или другом цветовом пространстве.

В данной работе представлен анализ различных методов, которые могут быть использованы при построении процедуры фильтрации, а также описан итоговый алгоритм, позволяющий решать поставленную задачу с приемлемым уровнем качества.

Разработка велась для повышения эффективности чтения (OCR) цветного текста на изменяющемся цветном фоне в интересах прикладной задачи считывания данных с удостоверяющих личность документов (паспортов, идентификационных карт, водительских удостоверений и др.).

Исходные данные. Постановка задачи

В распоряжении имеем набор цветных RGB-изображений области с текстом, вырезанной из отсканированных изображений документов. Некоторые примеры из серии таких изображений представлены на рисунке 1.

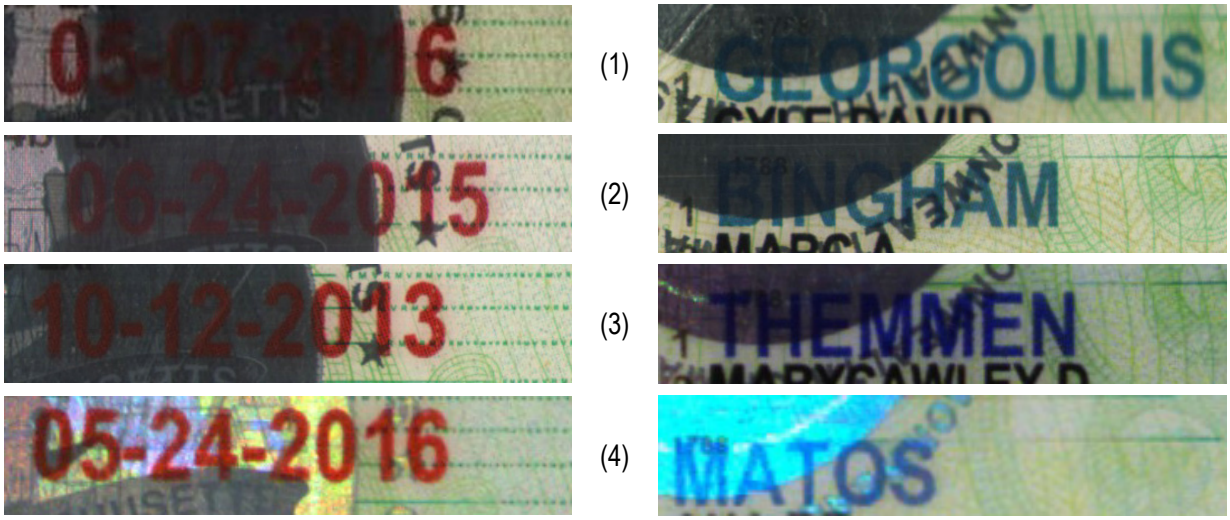


Рис 1. *Изображения текстовых полей, вырезанные из сканов идентификационных карт (а) и водительских удостоверений (б) штата Массачусетс (США).*

Процедура фильтрации предполагает предварительное назначение (вручную) точек текста и точек фона. Поэтому мы можем перенести рассматриваемую задачу в область теории распознавания образов [Журавлев, 1978]. В этом случае под множеством объектов X понимается множество всех пикселей изображения, которое необходимо разделить на подмножество пикселей текста X^{text} и подмножество пикселей фона X^{fond}. Каждый объект (пиксель) описывается 3-хмерным вектором признаков $x = (f_1,f_2,f_3)$. Значения компонент f_i вектора зависят от выбранной цветовой модели представления изображения, например в RGB-пространстве вектор признаков будет иметь вид x = (r,g,b), а для HLS модели x = (h,l,s). В качестве обучающей выборки будет выступать те пиксели, которые мы пометили как принадлежащие тексту $X^{m\ text} = \{x_1,x_2,\ldots x_m\}$ и пиксели, принадлежащих фону $X^{n\ fond} = \{x_1,x_2,\ldots x_n\}$.

Если для формализации понятия сходства между цветом произвольного пикселя изображения и цветом пикселей из обучающей выборки ввести функцию расстояния $\rho(x,x')$ то можно говорить о классической задаче метрической классификации [Айвазян, 1989]. Чем меньше значение этой функции, тем более схожи цвета двух сравниваемых пикселей x и x'.

К метрическим алгоритмам классификации относятся: метод ближайших соседей, метод потенциальных функций, метод парзеновского окна, алгоритм вычисления оценок и др. [Журавлев, 2006]. Фактически, выбор того или иного алгоритма классификации будет определять способ построения разделяющей классы поверхности.

Специфические особенности решаемой задачи накладывают ограничения на применимость тех или иных методов. Например, малый объем обучающей выборки $X^{m\ text}$ и $X^{n\ fond}$ (как правило, не более десяти точек для каждой подвыборки) не позволяет использовать методы байесовской классификации, в том числе метод парзеновского окна. Тот факт, что все включенные в обучающую выборку точки имеют одинаковую

важность, практически сводит метод потенциальных функций и алгоритм вычисления оценок к методу ближайших соседей. В дальнейшем мы проанализируем эффективность применения для решаемой задачи фильтрации метода ближайшего соседа (k=1), а также влияние на результат вида ядра функции расстояния $\rho(x,x')$.

Следующий важный аспект, вытекающий из анализа серии изображений, заключается в довольно большом разбросе цветовых характеристик как текста, так и фона. На изменчивость цвета фона и текста, оказывают влияние следующие факторы:

1) разброс параметров оптико-электронного тракта и параметров осветителей различных сканеров документов;
2) свечение голограммы (Рис.1, а(4) и б(4)) из-за изменения условий освещения или ориентации документа при сканировании;
3) изменение краски печати в различных партиях (редакциях) однотипных документов (Рис.1, б(1,2) и б(3,4));
4) зависимость цвета текста от цвета бланка, на котором он напечатан, обусловленная определенной степенью прозрачности краски, которой печатается текст.

Статистический анализ цветовых характеристик всего набора изображений позволяет сделать следующие выводы:

- в необработанных изображениях цвет текста сильно перемешан с цветом фона;
- факторы 2), 3) и 4) приводят к формированию кластеров цветовых характеристик текста с произвольным распределением точек внутри каждого кластера;
- фактор 1) приводит к размытию (расширению) кластеров в цветовом пространстве признаков;
- фактор 1) влияет глобально на всё изображение, т.е. почти синхронно смещает в цветовом пространстве не только цвет текста, но и цвет фона.

Для изображений текстовых областей (Рис. 1) данные выводы можно продемонстрировать следующей условной диаграммой. Для большей наглядности приведем цветовые характеристики текста в плоскости «цветовой тон - яркость» (Рис. 2).

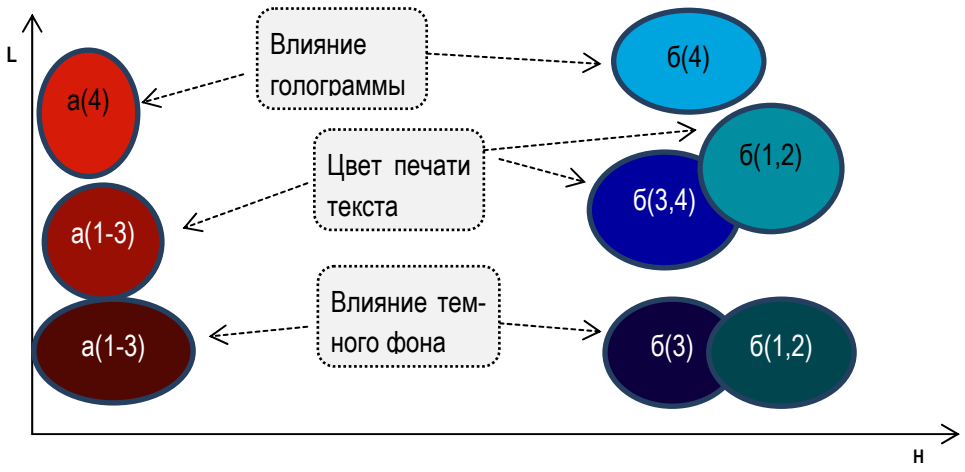


Рис. 2. Распределение цветовых характеристик текста в плоскости HL.

Очевидно, что перед фильтрацией необходимо попытаться устранить влияние фактора 1) с целью увеличения компактности кластеров цвета текста, а также стабилизации цветовых характеристик фона.

Таким образом, задача фильтрации цвета текста от цвета фона разделяется на две подзадачи: (I) компенсация аппаратных цветовых искажений изображения и (II) фильтрация – преобразование цвета из трехкомпонентного пространства в одномерное пространство полутонового изображения, где белым цветом будет обозначаться цвет фона, а черным – цвет текста.

Компенсация цветовых искажений

По условиям задачи для обработки предоставляются уже полученные с различных сканирующих устройств изображения. Поэтому нам недоступны такие процедуры нормализации изображений как калибровка видеотракта сканеров или проведение тестовых измерений их цветопередачи [Кровопусков, 2010]. Для разработки алгоритма компенсации цветовых искажений обратимся к методам представления цветовых характеристик, которые могут быть условно разделены на два класса: цветовые гистограммы [Гонсалес, 2006] и статистические модели представления цвета [Strieker, 1995].

Использование статистических моделей затруднительно в силу того, что

1) отсутствует нормальный закон распределения цветовых характеристик внутри каждого кластера (сложно для кластера определить «эталонное» значение цвета, к которому должны стягиваться цвета обрабатываемых изображений);
2) в то время как в обучающую выборку входят представители всех кластеров, то в отдельно взятом изображении присутствуют лишь некоторые из них, а иногда, что гораздо хуже, цвет фона может пересекаться с цветом имеющихся в выборке, но отсутствующих на изображении кластеров текста (в этом случае, без выполнения компенсации цвет фона будет принят за цвет текста);
3) для более качественного разделения текста и фона в обучающую выборку в качестве опорных могут включаться «граничные» цвета, множество которых на изображении, как правило, крайне мало.

Гистограммная форма описания изображения представляет собой информацию о распределении значений цветовых компонент в цветовых каналах. Для одного канала изображения I гистограмма является N-размерным вектором $H_I = (h_1, h_2, \ldots h_N)$, где h_i – отношение числа пикселей цвета i к общему числу пикселей в изображении I, а N – число квантов, например, градаций яркости.

В интересах компенсации аппаратных цветовых искажений мы применили процедуру пересчета значений цветовых компонент точек из обучающей выборки с учетом гистограмм обрабатываемого изображения и изображения, на котором данная точки была назначена. Перед обработкой (фильтрацией) изображения I для каждой m-ой точки из обучающей выборки значение k-ой цветовой компоненты f_k будет пересчитываться в соответствии с формулой:

$$f_k^{m\prime} = h_k^{I\,Min} + (f_k^m - h_k^{m\,Min}) \cdot \frac{(h_k^{I\,Max} - h_k^{I\,Min})}{(h_k^{m\,Max} - h_k^{m\,Min})}, \quad (1)$$

где $h_k^{I\,Min}$ и $h_k^{I\,Max}$ означают левую и правую границы гистограммы k-го канала изображения I, а $h_k^{m\,Min}$ и $h_k^{m\,Max}$ – границы гистограммы k-го канала того изображения, на котором была назначена m-я точка обучающей выборки.

Выражение (1) обеспечивает независимую для каждого канала подстройку значений опорных точек обучающей выборки под цветовые характеристики обрабатываемого изображения. Это очень важный момент, так как только в этом случае мы получаем возможность компенсировать различную цветопередачу сканирующих устройств. Процедуры типа автоконтрастирования для этих целей

непригодны, так как они выполняют пропорциональный для всех каналов пересчет значений. В результате, как правило, еще больше увеличивается расстояние между цветом опорных точек и цветом объекта, которому эти точки должны соответствовать (Рис. 3).

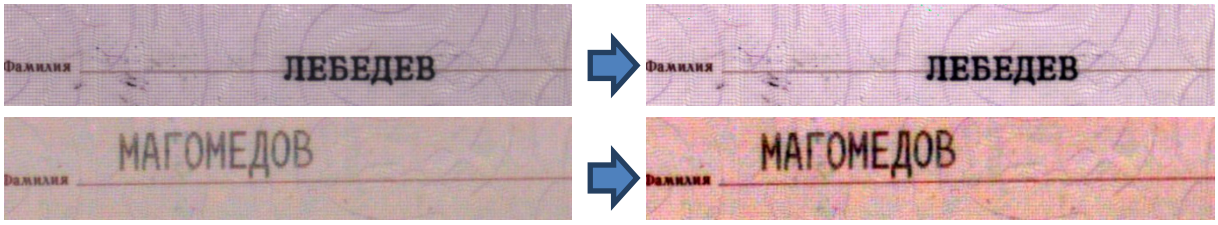


Рис. 3. Пример увеличения цветового рассогласования после применения процедуры автоконтрастирования.

Процедура фильтрации

Этот этап предполагает непосредственное преобразование цвета из трехкомпонентного пространства в одномерное пространство (полутоновое изображение). Значение яркости получаемого изображения должно быть пропорционально степени отличия цвета пикселя от цвета точек из множества $X^{m\ text}$ и степени сходства с цветом точек из множества $X^{n\ fond}$. Строго говоря, множество $X^{n\ fond}$ может быть и пустым. В большинстве случаев оказывается достаточным заполнить выборку только точками, задающими цвет текста.

Функция свертки трехкомпонентного значения цвета в значение яркости является не чем иным как функцией расстояния $\rho(x,x')$. При определении принципов работы функции расстояния мы исходили из следующих требований:

1) значение функции $\rho(x,x')$ должно зависеть от ближайшего соседа из выборок $X^{m\ text}$ и $X^{n\ fond}$;
2) функция $\rho(x,x')$ должна преобразовывать цветовое рассогласование $d(x,x')$ в значение яркости результирующего изображения.

Функция d(x,x') определяет ядро функции расстояния. В работе мы исследовали следующие ядра:

манхеттенское расстояние (норма L_1)

$$d(x,x') = \sum_{i=1}^{3} \left| x_i - x_i' \right|, \tag{2}$$

евклидово расстояние (норма L_2)

$$d(x,x') = \sum_{i=1}^{3} \left(x_i - x_i' \right)^2, \tag{3}$$

максимум модулей

$$d(x,x') = \max_i \left| x_i - x_i' \right|. \tag{4}$$

Преобразование отклика $d(x,x')$ в значение яркости осуществлялось посредством функции, имеющей следующий вид:

$$\rho(d) = \begin{cases} 0, \text{ если } d \le T/4 \\ \dfrac{255 \cdot (d - T/4)}{3 \cdot T/4}, \text{ если } T/4 < d < T \\ 255, \text{ если } d \ge T \end{cases}, \quad (5)$$

где T – параметр, определяющий толерантность или обобщающую способность функции расстояния.

В силу того, что значение функции d(x,x') вычисляется по расстоянию до ближайшего соседа, метод ближайшего соседа фактически реализует линейный классификатор, который определяет границу между классами как кусочно-линейную поверхность в цветовом пространстве признаков [Журавлев, 2006]. Вычисляемое в соответствии с (5) значение яркости можно интерпретировать как степень схожести цвета пикселя с заданным выборкой $X^{m\ text}$ цветом текста.

Функция расстояния (5) обеспечивает построение разделяющей поверхности даже при условии отсутствия в выборке точек фона, используя значение толерантности как максимально удаленную границу, отделяющую текст от фона. Если множество $X^{n\ fond}$ непустое и находится такое расстояние между x_i^{text} и x_j^{fond}, что $d(x_i^{text}, x_j^{fond}) < T$, то это означает, что j-ая опорная точка фона ограничивает область i-ой опорной точки текста. Для изображения б(4), Рис 1 сказанное можно проиллюстрировать следующей диаграммой (Рис. 4):

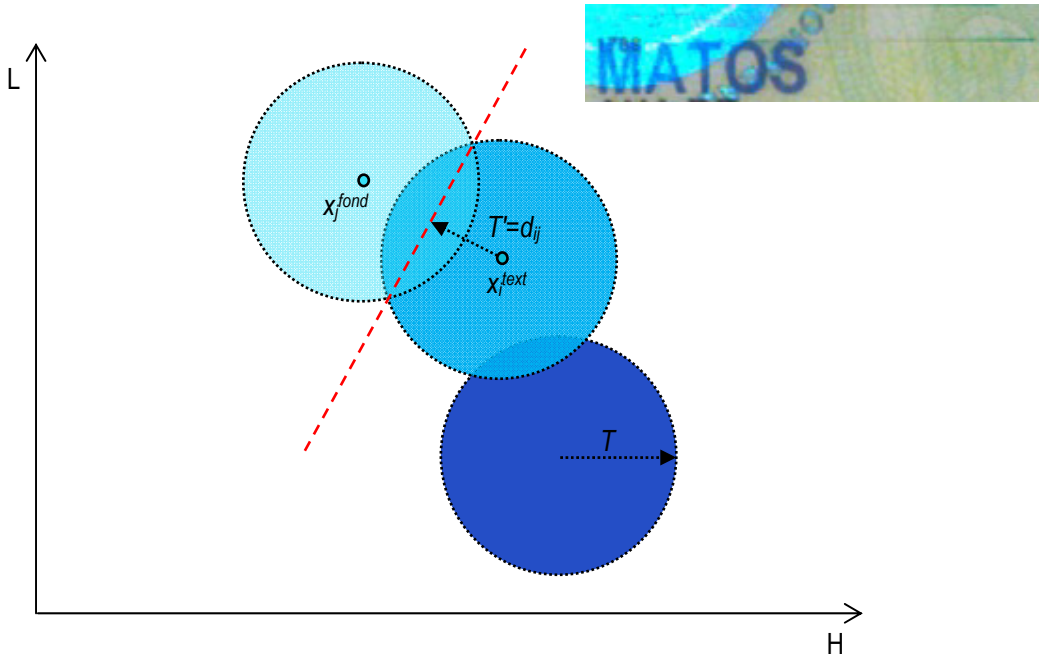


Рис. 4. Построение разделяющей поверхности в плоскости HL.

Из рисунка 4 и следует вывод, что нет необходимости задаваться опорными точками фона в случае, когда цвет фона отличается от цвета текста более чем на заданную величину толерантности T.

Выбор рабочей цветовой модели

Принятие решения о том, в какой цветовой модели осуществлять фильтрацию должно основываться на выполнении двух условий: во-первых, наилучшая цветовая модель должна увеличивать расстояние

между точками объекта и фона, и во-вторых, множество точек объекта должно располагаться как можно компактнее в цветовом пространстве выбранной модели.

Мы выработали следующие рекомендации:

1) если цвет текста или цвет фона можно отнести к категории черного, белого или серого, т.е. без явно выраженного цветового тона, то обработку следует осуществлять в RGB-пространстве;
2) если же цвет текста и фона можно назвать «цветными», то более эффективным будет использование перцепционной цветовой модели (HLS, HSV и др.).

Между HLS и HSV мы отдаем предпочтение первой, так как насыщенность в модели HLS всегда изменяется от полностью насыщенного цвета к эквивалентному серому цвету, в то время как в модели HSV при V=1 полностью насыщенный цвет переходит к белому.

Использование для обработки так называемых равноконтрастных цветовых моделей (Lab, LCH и др.) также возможно для ситуации 2), но будет оправдано только в тех случаях, когда необходимо алгоритмическую оценку цветоразличия привести в соответствие с человеческим цветовосприятием [Fairchild, 2005]. Платой за это станут повышенные вычислительные затраты. Использование равноконтрастных моделей имеет принципиальное значение, например, для кодирования изображений, как части графической системы, но не является обязательным в решаемой нами задаче.

В нашей работе мы использовали модифицированную HLS-модель: компоненту насыщенности S мы заменили на хрому C (Chroma) или ненормированную чистоту цвета, которая отражает степень приближения данного цвета к чистому спектральному цвету. В модели HLC хрома вычисляется по выражению

$$C = M - m,$$

где M = max(R,G,B) *и* m = min(R,G,B).

В отличие от насыщенности S значения хромы C не растянуты в диапазон [0,1]. Замена S на C преобразует цилиндрическое цветовое пространство HLS в форму трехмерного веретена. Хрома C максимальна только на средних уровнях яркости L, при увеличении или уменьшении яркости ощущение насыщенности падает (Рис. 5). Таким образом, при расчете цветового рассогласования, хрома будет вносить зависящий от уровня яркости вклад в итоговое расстояние.

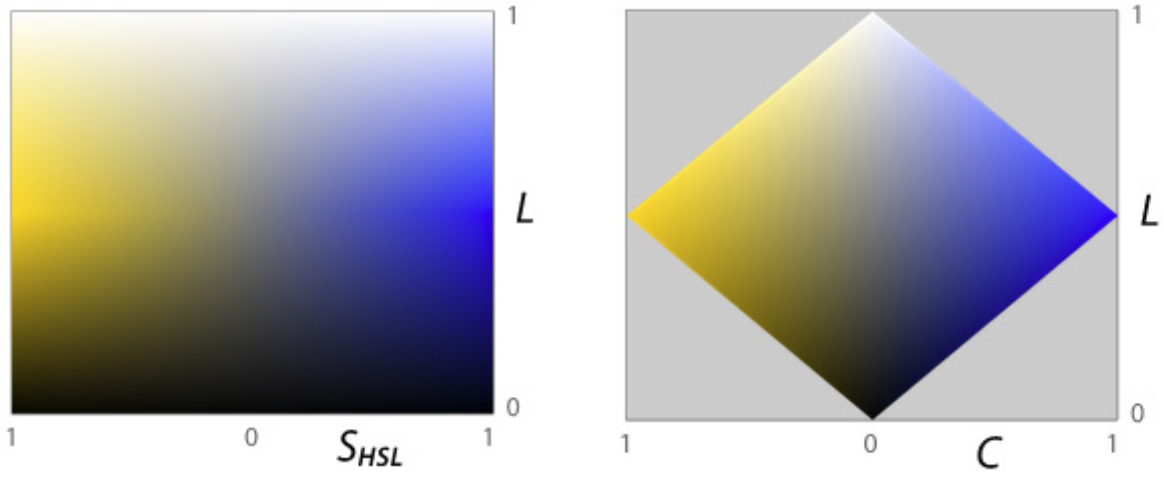


Рис. 5. Области определения насыщенности S в модели HLS и хромы C в модели HLC.

Результаты экспериментов

В качестве критерия эффективности процедуры фильтрации мы выбрали косвенный показатель – среднюю вероятность распознавания отфильтрованных изображений символов. Выбор данного критерия не противоречит здравому смыслу, так как обеспечивает оценку степени приближения формы символов к их эталонным изображениям. В тоже время такой способ оценки освобождает нас от весьма трудоемкого анализа правильности классификации каждого пикселя изображения.

Создание выборки опорных точек текста, а при необходимости и фона выполнялось вручную на нескольких изображениях с визуальной оценкой качества выполнения фильтрации. Здесь же выставлялось и значение толерантности T. Назначалось не более 8-ми точек в каждом подмножестве $X^{m\ text}$ и $X^{n\ fond}$. Затем процедура фильтрации запускалась на всем множестве изображений (от 500 до 1000) и подсчитывалась средняя вероятность распознавания символов отфильтрованного текста.

Настройка процедуры фильтрации, а также ее тестирование осуществлялись с применением различных ядер функции расстояния: dL1 (2), dL2 (3) и dmax (4). Кроме того, использовались два вида представления изображений: в цветовом пространстве RGB и в пространстве HLC. Примеры результатов фильтрации изображений (см. Рис. 1) представлены на рисунках 6 и 7.

Рис 6. Результат фильтрации текстового поля идентификационной карты штата Массачусетс (ID Mass).

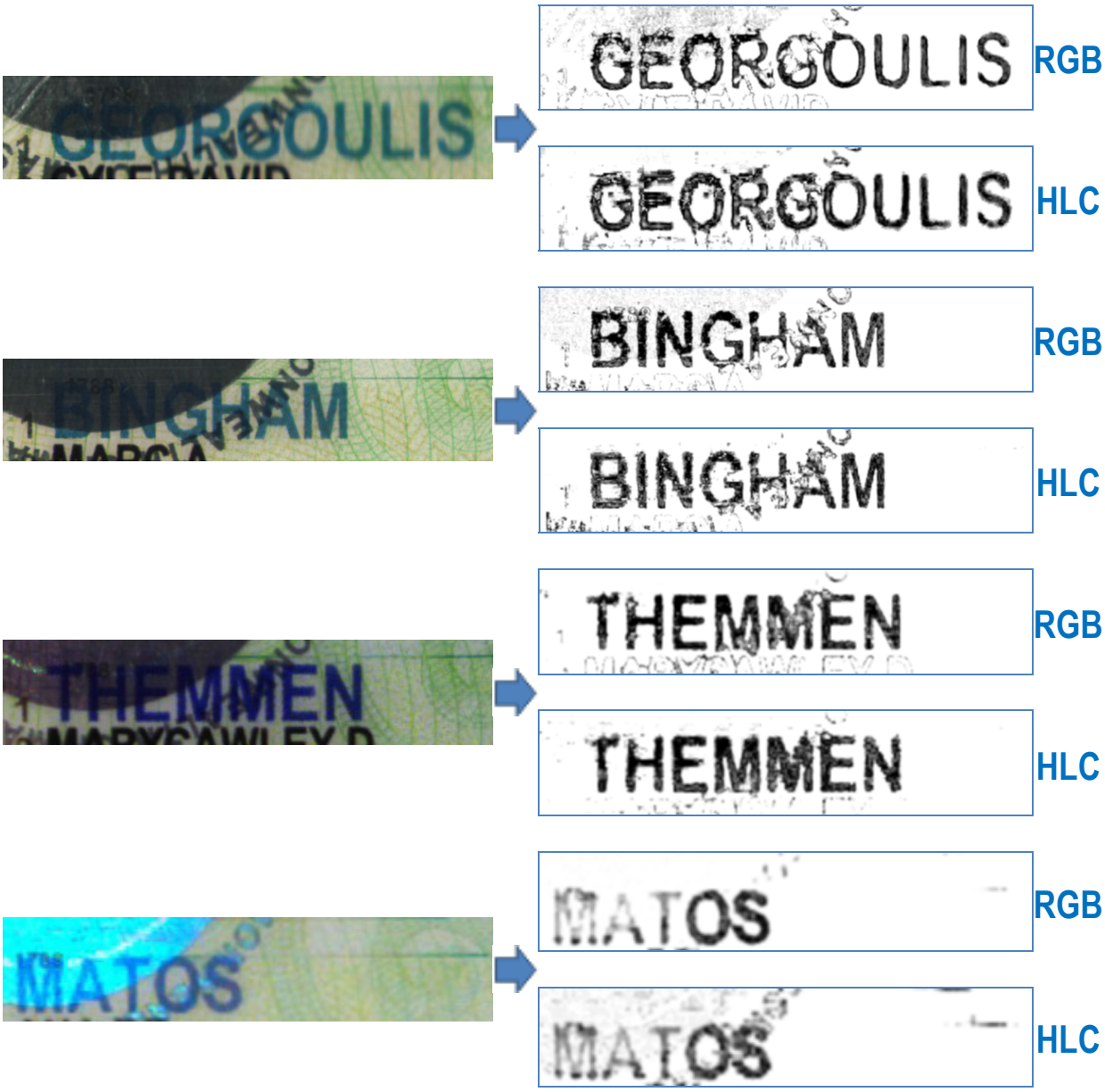


Рис 7. Результат фильтрации текстового поля водительского удостоверения штата Массачусетс (DL Mass).

Кроме указанных на рисунках 6 и 7 документах тестирование проводилось на текстовых полях еще трех видов документов. Примеры изображений текстовых полей приведены на Рис. 8.

Итоговые результаты тестирования процедуры фильтрации сведены в таблицы 1 и 2.

В последней строке таблицы под обозначением Def приводятся значения вероятностей распознавания изображений символов, обработанных с использованием таких инструментов, как выбор канала или смеси 2-х каналов, контрастирование и применение низкочастотных и высокочастотных фильтров.

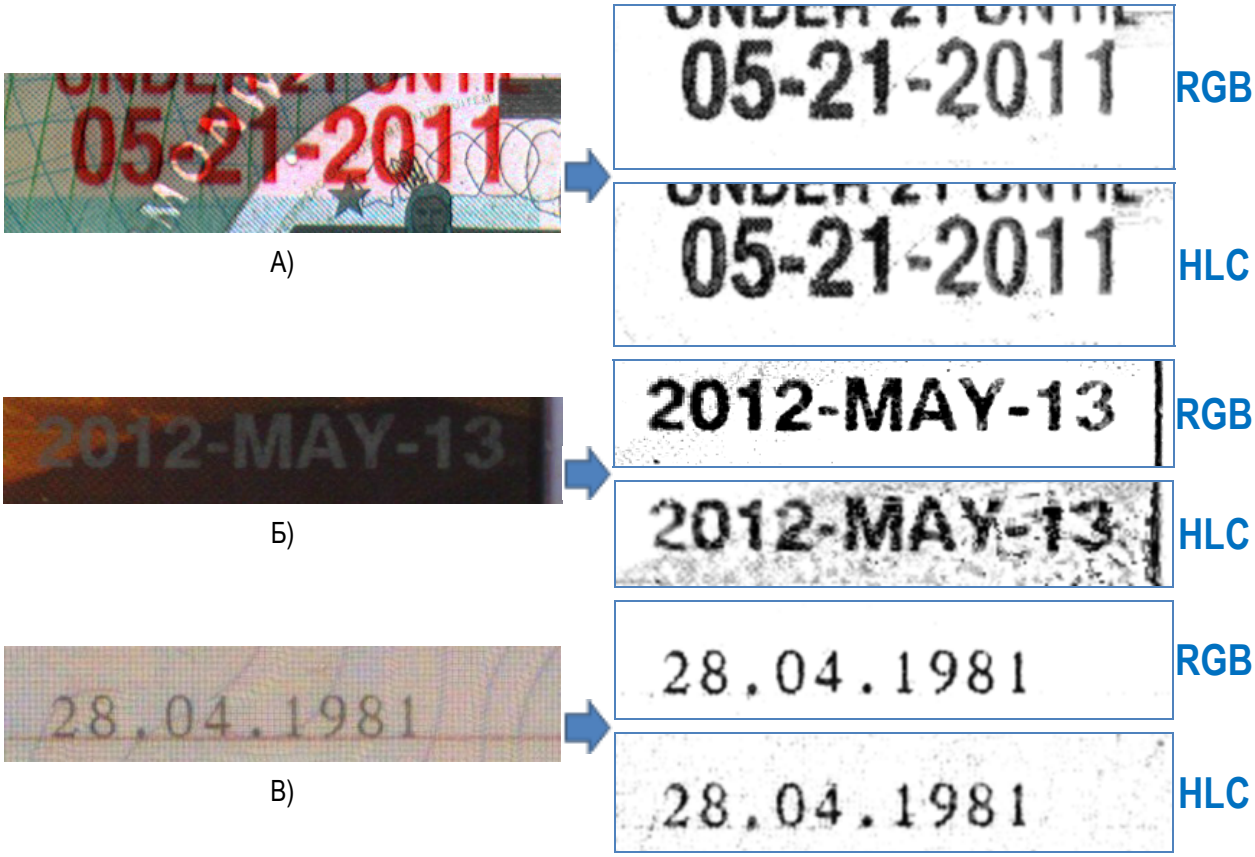


Рис 8. Изображения текстовых полей: а) водительское удостоверение для несовершеннолетних штата Массачусетс (DL Mass 18); б) водительское удостоверение штата Британская Колумбия, США (DL BC); в) Российский национальный паспорт (Pass RU).

Таблица 1. Средняя вероятность распознавания с применением RGB-модели.

Ядро функции сравнения	Документ 1 (ID Mass)	Документ 2 (DL Mass)	Документ 3 (DL Mass 18)	Документ 4 (DL BC)	Документ 5 (Pass RU)	Среднее по документам
d_{L1}	0,94	0,86	0,95	0,79	0,92	0,892
d_{L2}	0,93	0,89	0,96	0,82	0,92	0,904
d_{max}	0,90	0,88	0,89	0,71	0,90	0,856
Def	0,78	0,75	0,81	0,64	0,86	0,768

Таблица 2. Средняя вероятность распознавания с применением HLC-модели.

Ядро функции сравнения	Документ 1 (ID Mass)	Документ 2 (DL Mass)	Документ 3 (DL Mass 18)	Документ 4 (DL BC)	Документ 5 (Pass RU)	Среднее по документам
d_{L1}	0,90	0,85	0,97	0,70	0,89	0,862
d_{L2}	0,92	0,86	0,96	0,72	0,91	0,874
d_{max}	0,88	0,85	0,91	0,60	0,90	0,828
Def	0,75	0,71	0,85	0,61	0,83	0,75

Наилучшие результаты при обработке изображений продемонстрировала метрика L_2 (евклидово расстояние). Почти с такой же эффективностью работает и сравнение цвета по модулю рассогласования (метрика L_1). Худшие результаты, полученные на ядре d_{max} объясняются тем, что в сравнении участвует только одна цветовая компонента, причем на соседних пикселях нередко наблюдался эффект переключения процедуры сравнения с одного компонента на другой.

Как было изложено в предыдущем разделе, обработка изображения в цветовой модели HLC будет обоснована в тех случаях, когда цвет текста и цвет фона являются яркими и насыщенными. Только в этом случае (Документ 3 – DL Mass 18) эффективность фильтрации в цветовом пространстве HLC превысила эффективность фильтрации в пространстве RGB.

Заключение

В работе исследовано решение прикладной задачи выделения цветного текста на сложном цветном фоне. Для повышения эффективности решения задачи предложен способ компенсации цветовых искажений, вызванных сканирующей аппаратурой. На базе метода линейной классификации разработана процедура фильтрации фона от текста путем свертки цветного изображения в полутоновое изображение.

Для практического применения разработанной процедуры сделаны конкретные рекомендации по выбору рабочего цветового пространства и ядра функции расстояния.

Экспериментально получено подтверждение эффективности применения разработанного фильтра в задаче OCR. По сравнению с традиционно используемыми средствами предварительной обработки применение фильтра позволило повысить вероятность распознавания символов на 6-18%.

Благодарности

Статья публикуется при частичной поддержке проекта ITHEA XXI Международного научного общества ITHEA (www.ithea.org) и Ассоциации Создателей и Пользователей Интеллектуальных Систем ADUIS (www.aduis.com.ua).

Литература

[Fairchild, 2005] Mark D. Fairchild. Color Appearance Models, 2nd Edition. John Wiley & Sons, 2005. – 408 p.

[Strieker, 1995] Strieker M., Orengo M. Similarity of color images // Storage and Retrieval for Image and Video Databases (SPIE). – 1995. – P. 381-392.

[Wikipedia, 2013] http://en.wikipedia.org/wiki/Comparison_of_raster_graphics_editors.

[Айвазян, 1989] Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: классификация и снижение размерности. – М.: Финансы и статистика, 1989.

[Гонсалес, 2006] Гонсалес Р., Вудс Р. Цифровая Обработка Изображений. Техносфера.Москва, 2006 - 432с.

[Журавлев, 1978] Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания и классификации // Проблемы кибернетики. – М.: Наука, 1978, вып. 33. – С. 5-68.

[Журавлев, 2006] Журавлев Ю.И., Рязанов В.В., Сенько О.В. «Распознавание». Математические методы. Программная система. Практические применения. – М.: Фазис, 2006.

[Кровопусков, 2010] Кровопусков А.В., Крыловецкий А.А., Рындин А.А. Метод устранения цветового рассогласования в системе активного трехмерного сканирования. – Вестник ВГУ. сер.: Физика. Математика, 2010. №2, С. 247-251.

Информация об авторах

Телятников Роман *– к.т.н., научный руководитель разработками ПО, ООО «Регула», 220036 ул. Волоха 1-314, г. Минск, Беларусь, e-mail: raman.tsialiatnikau@regula.by*

Основные направления деятельности: распознавание образов, обработка изображений, нейрофизиология

Шумский Иван *– к.т.н., директор ООО «Регула», 220036 ул. Волоха 1-314, г. Минск, Беларусь, e-mail: ivan.shumsky@regula.by*

Основные направления деятельности: автоматизация анализа подлинности документов и банкнот: проектирование оборудования и программного обеспечения

Мамедов Ариф *– к.х.н., президент Regula Forensics, Inc., 1800 Alexander Bell Drive, Suite 400 Reston, VA 20191, USA, e-mail: arif.mamedov@regula.us*

Основные направления деятельности: подлинность документов и банкнот: маркетинговые исследования, автоматизация и проектирование

Протосавицкий Анатолий *– инженер-программист, ООО «Регула», 220036 ул. Волоха 1-314, г. Минск, Беларусь, e-mail: anatol.pratasavitski@regula.by*

Основные направления деятельности: обработка цветных изображений, элементы защиты и системы проверки подлинности документов

Матусевич Екатерина *– инженер-программист, ООО «Регула», 220036 ул. Волоха 1-314, г. Минск, Беларусь, e-mail: katsiaryna.harshkova@regula.by*

Основные направления деятельности: обработка цветных изображений, элементы защиты и системы проверки подлинности документов

Степанькова Екатерина *– инженер-программист, ООО «Регула», 220036 ул. Волоха 1-314, г. Минск, Беларусь, e-mail: KVayavodava@regula.by*

Основные направления деятельности: обработка изображений, статистический анализ, проверка подлинности документов

О ВАРИАТИВНОСТИ НЕКОТОРЫХ БУКВЕННЫХ ЧАСТОТ В СУПРАСЛСКОМ СБОРНИКЕ

Й. Табов, Св. Христова

Абстракт. *В настоящей работе предлагаем начальные шаги исследования вариативности частот некоторых букв в текстах разной длины из Супраслского сборника. Получены интервалы, в которых варьируют частоты букв М, Н, Е и ѣ в текстах длиной одной до пяти страниц сборника.*

Keywords*: буквенные частоты, старославянские тексты, Супраслский сборник*

Введение

Среди важнейших задач при изучении болгарского письменного наследия можно отметить атрибуцию и датирование старых болгарских текстов — определение автора текста, школы, к которой его причисляют, и время его создания. Подобной проблемой является установление однородности текста (наличие вставок другого автора или переписчика). Эту традиционную для палеографов задачу не всегда можно успешно решить только их усилиями — сравнением специфических характеристик текста. Желательно найти формальный, количественно-статистический подход, основанный на анализе буквенных частот, который давал бы дополнительные доводы, дающие предпочтение одной палеографской гипотезы перед остальными.

Для создания такого подхода нужны масштабные исследования старых болгарских текстов, одним из важнейших среди которых является Супраслский сборник. Он подходит для экспериментов с буквенными частотами еще и потому, что результаты от них можно сравнивать с уже имеющимися многочисленными палеографическими анализами.

В настоящей работе предлагаем начальные шаги таких исследваний: сравнения частот некоторых букв в текстах разной длины из Супраслского сборника по его факсимильному изданию [Супрасълски сборник, 1983].

1 О вариативности буквы М

Диаграмма на **рис. 1** показывает наглядно распределение частот буквы М для 37 страниц рукописи Супраслского сборника: по абсциссе частоты в процентах, разделенные на интервалы (от 1% до 1,5 %, от 1,5 % до 2 % и т.д.), а по ординате – число страниц, на которых частоты на буквы М попадают в соответствующий интервал, отмеченный на абсциссе.

Диаграмма на **рис. 2** показывает наглядно распределение частот буквы М для 36 пар страниц (36 текстов, занимающих 2 страницы) в рукописи Супраслского сборника: по абсциссе частоты в процентах, разделенные на интервалы (от 1% до 1,5 %, от 1,5 % до 2 % и т.д.), а по ординате – число страниц, для которых частоты буквы М попадают в соответствующий интервал, отмеченный на абсциссе.

Аналогичная диаграмма для 35 троек страниц представлена на **рис. 3**.

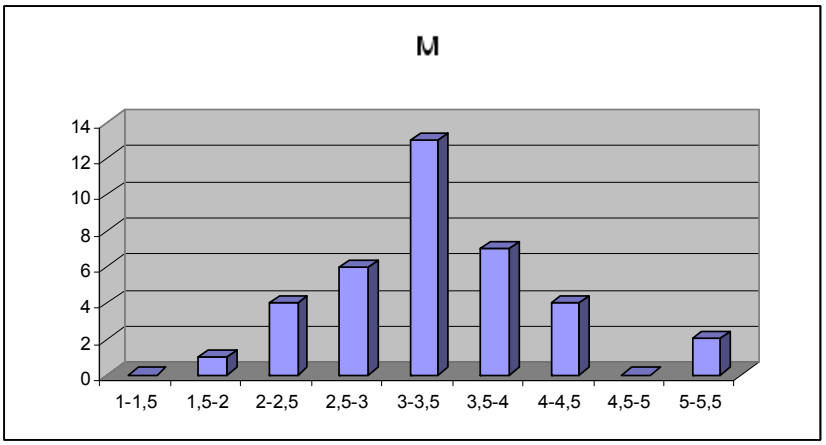


Рис. 1. Распределение частот буквы М для 37 страниц рукописи Супраслского сборника: по абсциссе частоты в процентах, разделенные на интервалы (от 1% до 1,5 %, от 1,5 % до 2 % и т.д.), а по ординате – число страниц, на которых частоты на буквы М попадают в соответствующий интервал, отмеченный на абсциссе.

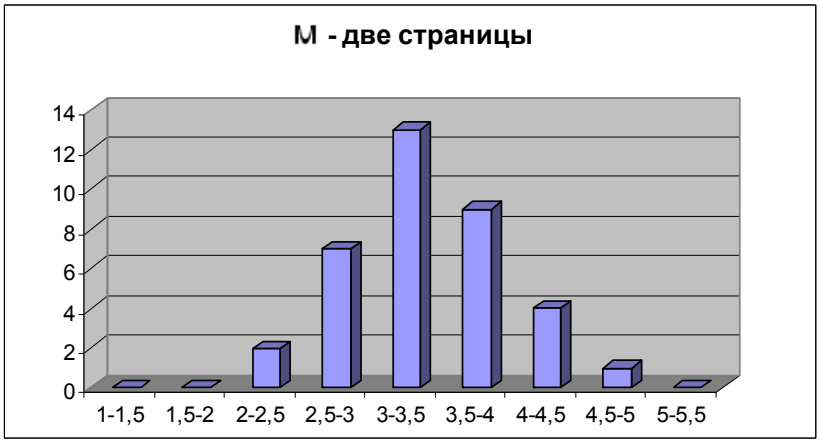


Рис. 2. Распределение частот буквы М для 36 пар страниц (36 текстов, занимающих 2 страницы) в рукописи Супраслского сборника: по абсциссе частоты в процентах, разделенные на интервалы (от 1% до 1,5 %, от 1,5 % до 2 % и т.д.), а по ординате – число пар страниц, для которых частоты буквы М попадают в соответствующий интервал, отмеченный на абсциссе.

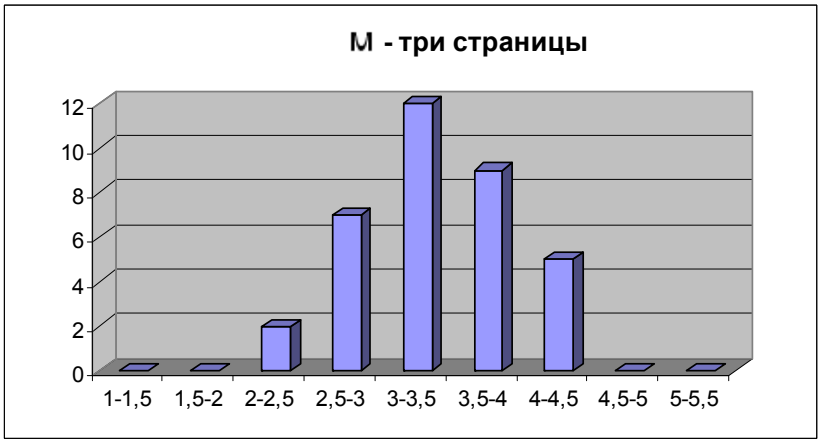


Рис. 3. Распределение частот буквы М для 35 троек страниц (35 текстов, занимающих 3 страницы) в рукописи Супраслского сборника: по абсциссе частоты в процентах, разделенные на интервалы (от 1% до 1,5 %, от 1,5 % до 2 % и т.д.), а по ординате – число троек страниц, для которых частоты буквы М попадают в соответствующий интервал, отмеченный на абсциссе.

В охваченных настоящим исследованием 37 страниц Супраслского сборника частота буквы М немного выше средней: около 3,343 %. Как видно из диаграммы на **рис. 1**, для этих страниц она изменяется от 1,5 % до 5,5 %, т.е. – в процентах – в интервале (1,5; 5,5). Для пар и троек страниц – судя по диаграммам на **рис. 2** и **рис. 3** - этот интервал сужается соответственно до (2 ; 5) и (2 ; 4,5).

Эти результаты дают возможность сравнить вариативность частот при уменьшении длины текстов: с троек к парам страниц и с троек к отдельным страницам (**рис. 4**)

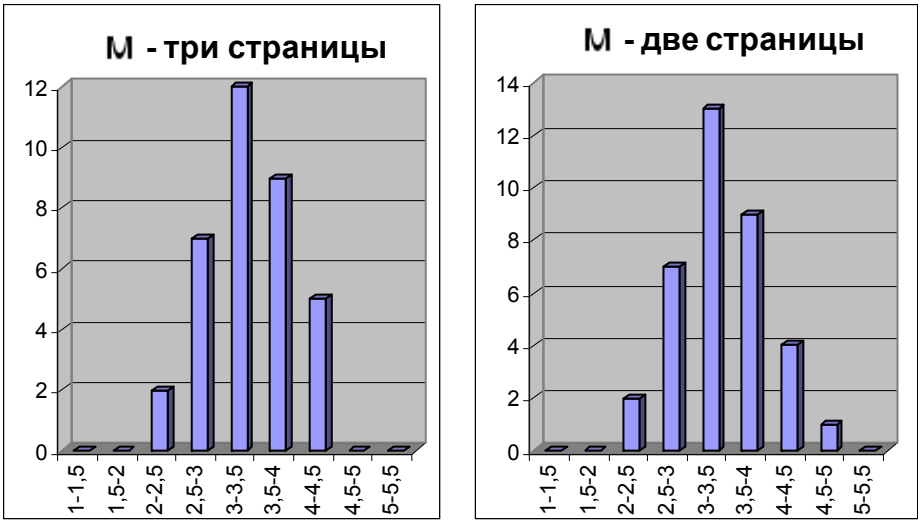


Рис. 4. Сравнение вариативности частот буквы М при уменьшении длины текстов: с троек (слева) к парам (справа) страниц в рукописи Супраслского сборника. Частоты М варируют соответственно в интервалах (2 ; 4,5) и (2 ; 5).

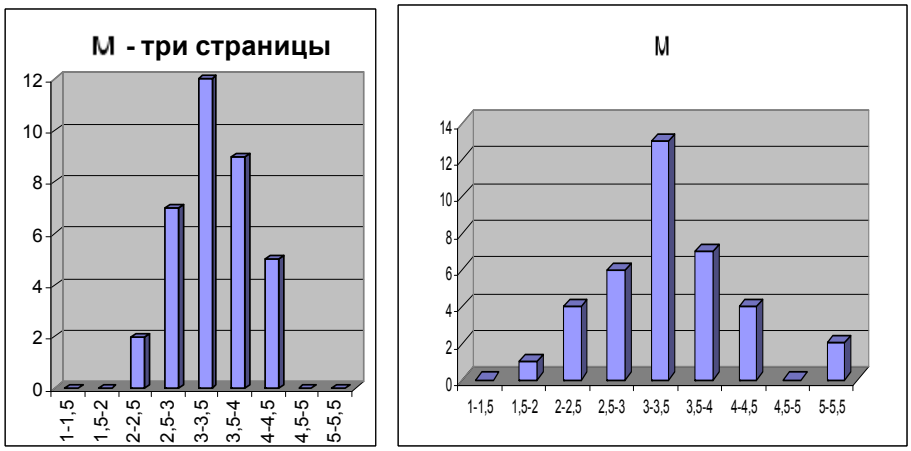


Рис. 5. Сравнение вариативности частот буквы М при уменьшении длины текстов: с троек (слева) к отдельным страницам (справа) в рукописи Супраслского сборника. Частоты М варируют соответственно в интервалах (2 ; 4,5) и (2 ; 5).

Диаграммы на **рис. 4** и **рис. 5** иллюстрируют расширение интервала вариативности частот буквы М при уменьшении объема соответствующего текста – ожидаемое изменение. В процентном соотношении оно означает расширение интервала соответственно на 20 % и 60 %.

2 О вариативности буквы Н

В охваченных нашим исследованием 37 страниц Супраслского сборника частота буквы Н большая, выше средней: около 9,052 %. Как видно из диаграммы на **рис. 6**, для этих страниц она изменяется от

5,7 % до 12 %, т.е. – в процентах – в интервале (5,7 ; 12). Для троек и пятерок страниц – судя по диаграммам на **рис. 7** и **рис. 8** - этот интервал сужается соответственно до (7,8 ; 11,3) и (7,8 ; 10,8).

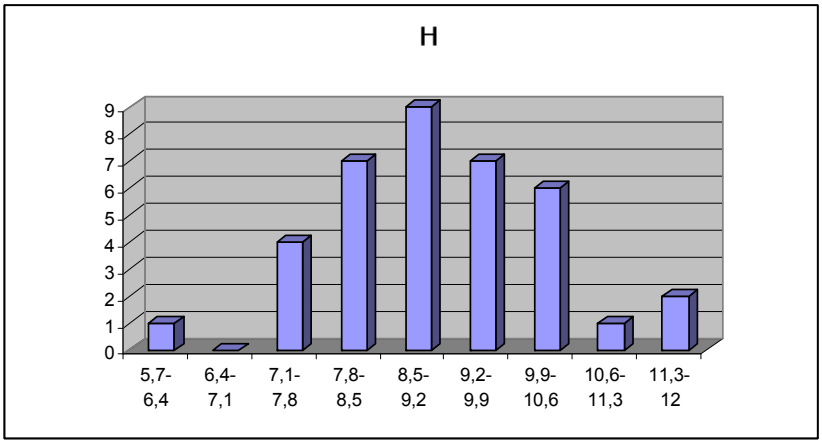


Рис. 6. Разпределение на частотите на буквата Н за 37 страници от ръкописа на Супраслския сборник: по абсцисата са частотите в проценти, а по ординатата – броят на страниците, в които частотите на буквата Н са в съответния интервал.

Эти результаты дают возможность сравнить вариативность частот при уменьшении длины текстов: с пятерок к тройкам страниц и с пятерок к отдельным страницам (**рис. 7** и **8**).

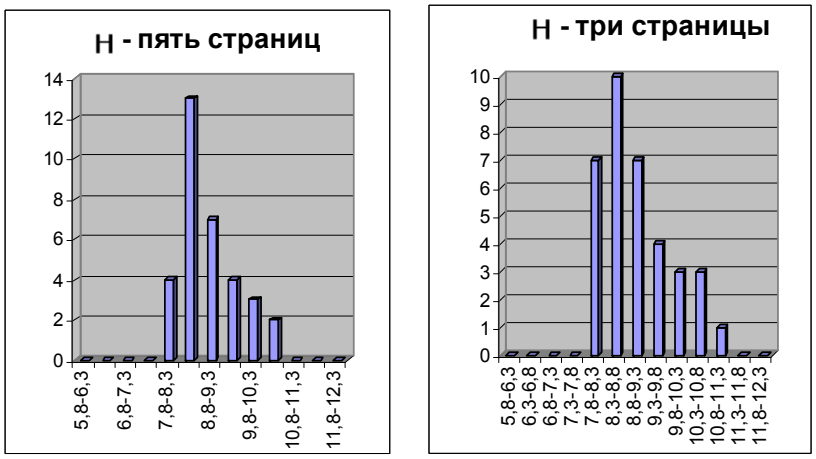


Рис. 7. Сравнение вариативности частот буквы Н при уменьшении длины текстов: с пятерок (слева) к тройкам страниц (справа) в рукописи Супраслского сборника. Частоты Н варируют соответственно в интервалах (7,8 ; 10,8) и (7,8 ; 11,3).

Диаграммы на **рис. 7** и **рис. 8** иллюстрируют расширение интервала вариативности частот буквы Н при уменьшении объема соответствующего текста – ожидаемое изменение. В процентном соотношении оно означает расширение интервала соответственно на 17 % и 110 %.

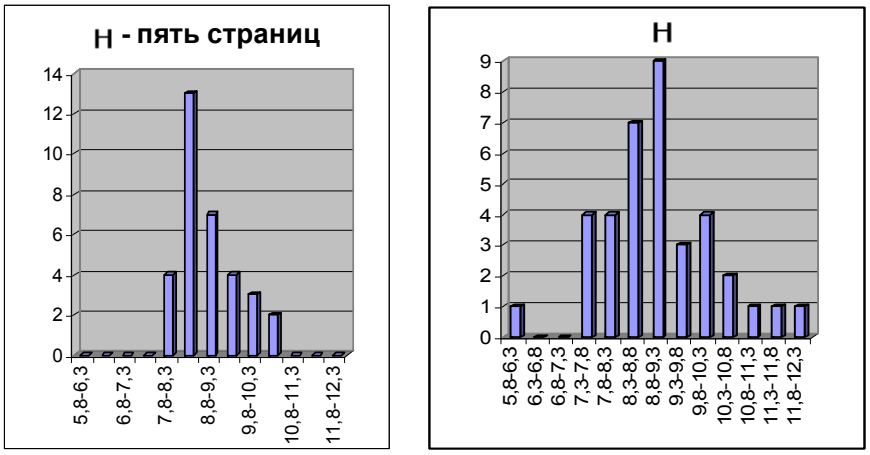


Рис. 8. Сравнение вариативности частот буквы Н при уменьшении длины текстов: с пятерок (слева) к отдельным страницам (справа) в рукописи Супраслского сборника. Частоты Н варируют соответственно в интервалах (7,8 ; 10,8) и (5,8 ; 12,3).

3 О вариативности буквы Е

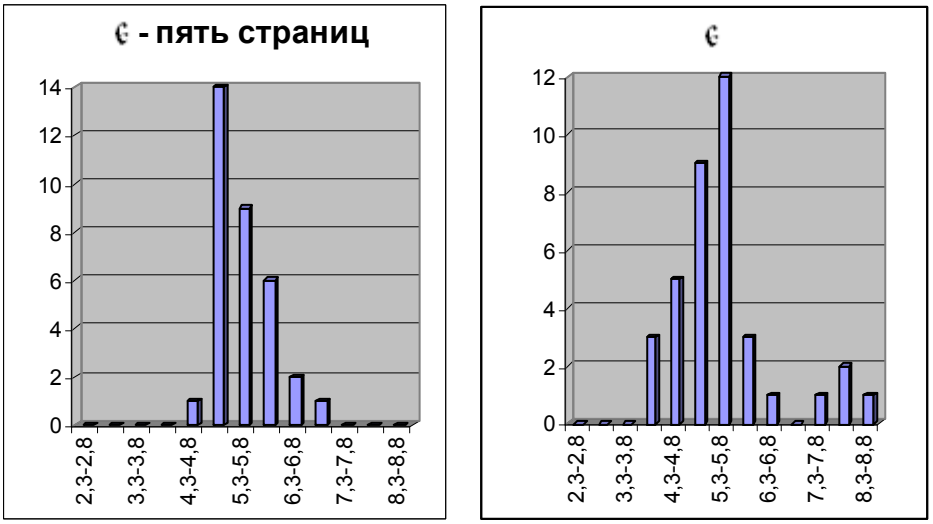


Рис. 9. Сравнение вариативности частот буквы Е при уменьшении длины текстов: с пятерок (слева) к отдельным страницам (справа) в рукописи Супраслского сборника. Частоты Е варируют соответственно в интервалах (4,3 ; 7,3) и (3,8 ; 8,8).

Диаграмма на **рис. 9** иллюстрирует расширение интервала вариативности частот буквы Е при уменьшении объема соответствующего текста: расширение интервала на 67 %.

4 О вариативности буквы Ѣ

Диаграмма на **рис. 10** иллюстрирует расширение интервала вариативности частот буквы Ѣ при уменьшении объема соответствующего текста: расширение интервала на 160 %.

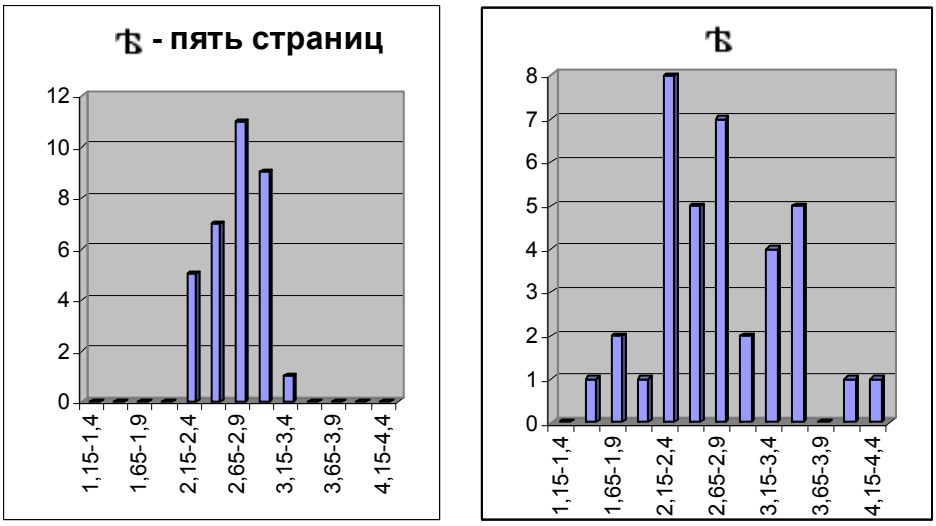


Рис. 10. Сравнение вариативности частот буквы ѣ при уменьшении длины текстов: с пятерок (слева) к отдельным страницам (справа) в рукописи Супраслского сборника. Частоты варируют соответственно в интервалах (2,15 ; 3,4) и (1,4 ; 4,4).

5 Выводы и перспективы

Число текстов фиксированной длины, для которых в нашем исследовании вычисляются частоты букв - от 33 до 37, что представляет собой выборку, достаточную для того, чтобы играть роль первоначального ориентира для выводов, относящихся к всему сборнику.

Судя по полученным нами данным, для текстов длиной в одну страницу частота данной буквы может варировать в интервале длиной до около 5 % (т.е. плюс-минус 2,5 %). Для текстов длиной в 5 страниц этот интервал значительно уже – около 3 %.

Предположим, что нужно сравнить два длинных текста на старославянском языке (длиной больше 5 страниц по масштабам Супраслского сборника) X и У, и что частоты некоторой буквы (хотя бы одной) в X и в У отличаются больше, чем на 1,5 %; это отличие указывало бы на разное происхождение этих текстов – что они писаны не одним и тем же человеком («одной и той же рукой»). Более существенные отличия буквенных частот указывали бы на то, что X и У являются творением разных школ.

Для того, чтобы установить точные параметры вариативности отдельных букв и использовать их на практике, нужны дальнейшие исследования

5 Библиография

[Супрасълски сборник, 1983] Супрасълски или Ретков сборник. Издат. на БАН, София, 1983.

Информация об авторах

Йордан Табов *– Институт математики и информатики БАН, ул. Акад. Г. Бончев блок 8, 1113 София, Болгария; e-mail: tabov@math.bas.bg*

Основные области научных исследований: Применения математики и информатики в гуманитарных науках, дидактика математики и информатики

МОДЕЛИ ИНТЕЛЛЕКТУАЛЬНОЙ АДАПТИВНОЙ ПОДДЕРЖКИ НАВИГАЦИИ В КОМПЬЮТЕРНЫХ ОБУЧАЮЩИХ СИСТЕМАХ

Игорь Шубин, Владимир Чернов, Владимир Гриценко, Ирина Кириченко

***Abstract**: The various questions of creation of integrated development environment for computer training systems are considered in the given paper. The information technologies that can be used for creation of the integrated development environment are described. The different didactic aspects of realization of such systems are analyzed. The ways to improve the efficiency and quality of learning process with computer training systems for distance education are pointed*

***Keywords**: Distance Learning, Learning Course Model, Computer-Based Training System, Hypermedia, Web-Based Design.*

Введение

В настоящее время обучение через Интернет является актуальной областью исследований и разработок. Польза от такого подхода в обучении очевидна: независимость расположения обучаемых и независимость от платформы. Приложение, установленное и поддерживаемое в одном месте, может использоваться тысячами обучаемых по всему миру, имеющих компьютер с любым видом подключения к Сети. Тысячи Web-курсов и других обучающих приложений стали доступны в последнее время.

Проблема заключается в том, что большинство из них являются ничем иным, как сетью статичных гипертекстовых страниц. Перспективной целью является разработка передовых образовательных приложений, основанных на Web, которые смогут предложить нечто значительное в плане интерактивности и адаптивности.

В последнее время адаптивные гипермедиа-системы становятся все более и более популярными в дистанционном обучении, предоставляя средства доступа к информации, управляемые пользователем. Адаптивные гипермедиа-системы сводят воедино идеи гипермедиа-систем и интеллектуальных обучающих систем и делают возможным персонализированный доступ к информации. Рассматриваются вопросы поддержки дистанционного обучения, особое внимание уделяется анализу методов и средств адаптивной гипермедиа, используемых современными адаптивными обучающими Web-системами.

Адаптация исключительно важна для образования с использованием глобальной сети, по меньшей мере, по двум основным причинам. Во-первых, большинство Web-приложений используются множеством таких различных пользователей, что не предполагалось при разработке локальных приложений и следовательно, Web-приложения, спроектированные для специфичного класса пользователей, не будут подходить другим пользователям. Во-вторых, во многих случаях пользователь работает один на один с Web-наставником или курсом [Bondarenko et al, 2008].

Таким образом, целью исследования является разработка методов и моделей построения систем адаптивной гипермедиа, основанной на упрощенном формате навигационных правил перемещения обучаемых по учебному материалу. Навигационные правила являются функциями алгебры конечных предикатов и предикатных операций, позволяющей описывать произвольные модели искусственного

интеллекта. Формат правил также является объектом разработки и должен позволять разработчикам описывать навигационные правила для их применения в автоматизированных обучающих системах. Кроме того, необходимо разработать инструментальное средство для проведения экспериментов с разработанными навигационными правилами с целью адаптации учебных материалов для студентов. В данном случае система адаптивной гипермедиа является адаптивным воздействием для предоставления учебной информации на основе знаний пользователя.

Главным отличием разрабатываемых математических моделей от аналогичных является то, что аналогичные системы, основанные на применении навигационных правил, не предоставляют разработчикам ни средств, ни функций, которые могут уменьшить сроки и трудоемкость составления правил.

Таким образом, в данной статье разработаны следующие компоненты адаптивной обучающей системы: модель процесса обучения, модель обучаемого, описание основной цели обучения, алгоритм обучения, разработан набор навигационных правил.

Обучающие гипермедийные адаптивные системы

В другом контексте, обучающие адаптивные системы в Web лишь одна из существующих разновидностей адаптивных систем. WWW показывает, что может являться хорошей платформой для разработки и тестирования различных адаптивных приложений. С одной стороны, это перспективно: системы в Web действительно нуждаются в адаптации, так как они работают с более значительно отличающимися пользователями, чем более ранние системы, предназначенные для установки непосредственно на машину пользователя. С другой стороны, Web дает комплексным адаптивным системам прекрасный шанс дотянуться до многих реальных пользователей. Пользователи Web также могут помочь разрешить проблемы оценивания, так как все данные о взаимодействии пользователей с адаптивной Web-системой могут быть записаны на централизованном сервере и использованы для обстоятельного анализа.

Системы адаптивной гипермедиа применяют различные виды моделей пользователя для адаптации контента автоматизированной обучающей системы (АОС) и ссылок внутри него под уровень знаний и интересы пользователя. Образование всегда было одной из главных областей применения адаптивной гипермедиа. Большинство адаптивных систем гипермедиа используют методы, которые позволяют разработчикам описывать навигационные правила перемещения обучаемых по контенту АОС [Shubin et al, 2011].

Методы построения интеллектуальной адаптивной поддержки обучения

Адаптивная поддержка (АП) совместной работы использует знания системы о различных пользователях (хранимых в моделях пользователя) для формирования сбалансированной группы для совместной работы. Технология адаптивной поддержки навигации помогает обучаемому в ориентации и навигации в гиперпространстве, изменяя появление видимых ссылок. В отдельных случаях система может адаптивно сортировать, аннотировать или частично прятать ссылки на текущей странице для облегчения выбора пользователем следующей ссылки. Адаптивная поддержка навигации (АПН) может рассматриваться как дополнение к адаптивному планированию в контексте гипермедиа. Она участвует в решение той же задачи – помочь обучаемому найти оптимальный путь в обучающем материале. В тоже время адаптивная поддержка навигации направляет обучаемого менее настойчиво, по сравнению с традиционным адаптивным планированием: она направляет обучаемого косвенным образом, разрешая выбрать следующий фрагмент заданий для изучения или следующую задачу для решения. Существует

несколько известных способов адаптации гиперссылок. Тремя наиболее популярными путями, используемыми в Сети, являются прямое руководство, применение адаптивной отметки ссылок и адаптивное сокрытие ссылок.

Разработка моделей адаптации в схемах обучения

Введем понятие «Схема обучения», как результат генерации специфического пути прохождения учебной дисциплины для обучаемого, которая включена в обучающую последовательность курса. В конце последовательности, взаимодействие пользователя оценивается тестированием на приобретенные знания. Согласно результатам теста, пользователь может переходить к следующему разделу обучающих материалов или проходить тот же курс, если необходимо модифицировать профиль обучаемого, учитывая пересданные материалы.

Проход от обучающего курса к другому не выполняется в произвольном или автоматическом виде, а согласно точным навигационным правилам. Эти навигационные правила необходимы для точного описания переходов в форме обучающей сети прохождения. Фактически, прохождение курса основано на наборе переходов между другими обучающими последовательностями курса, в развитии пути, пока последняя последовательность курса не будет достигнута. Сеть прохождения отображает зависимость учебных курсов, т.е. каждый учебный курс, может быть зависим от одного или нескольких учебных процессов. Это подразумевает, чтобы начать обучение по курсу, нужно обладать знаниями по зависимым к нему предыдущим курсам. Обучающая сеть прохождения представляет все навигационные правила, интерпретирующие различные обучающие последовательности для достижения образовательной цели. Второй этап предназначен для проверки теоретических знаний обучаемого. С этой целью используется один из тестов. Если у обучаемого есть ошибки, ему необходимо возвратиться и снова пройти теоретический курс. В общих чертах, в течение первых и вторых этапов, адаптация гипермедийной обучающей системы ориентирована на обучаемого и поддерживается выбором важных учебных материалов и тестов.

В течение третьего этапа, обучаемый решает проблемы по предмету под адаптивным управлением. В каждой мере изучения этого этапа, согласно результатам решения проблемы, принимается решение по:

1) продолжению изучения (обучаемому дано курс на требуемые ему знания);
2) успешное достижение цели (цель изучения достигнута);
3) окончание этапа (обучаемый направляется на изучение новой теории).

Необходимо должным образом подчеркивать и описывать все о путях этих компонентов и порядок компонентов, которые важны для обучаемого: обучаемый курс материалов, тест или внутренний путь, динамически отформатированые результаты решения проблем [Святкин, 2012].

Методы адаптации делятся на два вида: адаптацию содержания (контента) и адаптацию связей.

Адаптация содержания приспосабливает контент узла к характеристикам обучаемого. Адаптация связи приспосабливает связи узла (гиперссылки на другие узлы). В данном случае более подходящей будет адаптация связей, потому что нам необходимо направлять обучаемых через гиперпространство, образованное контентом АОС.

Адаптация связей делится на четыре типа согласно тому, как приспосабливаются связи: «прямое руководство» предполагает объяснение к связи, за которой пользователь должен следовать или создание кнопки для направления пользователя; «адаптивный заказ» основан на определении степени пригодности каждой связи для пользователя; «технология адаптивного сокрытия ссылок» сужает доступное гиперпространство, скрывая неадекватные обучаемому связи; «адаптивная аннотация» – предполагает

художественные дополнительные оформления, такие как изображения и цвета связей. Метод сокрытия связей не показывает связей, не адекватных текущим характеристикам обучаемого. Хотя это вынуждает пользователя следовать за навигационными путями созданными разработчиком АОС, сужая его свободу перемещения по контенту, данный метод позволяет разработчику направить его по наиболее оптимальному пути.

Для разработки интеллектуальных АОС предлагаемой архитектуры авторы выбрали для использования модель, хранящую как долгосрочную так и краткосрочную информацию об обучаемом. Для того чтобы модель обучаемого была более простой упростим формат навигационного правила. В качестве параметра обучаемого будем использовать интересы и уровень знаний пользователя. Это необходимо для моделирования стратегии обучения в течение продолжительного времени.

Также будем использовать историю взаимодействия обучаемого с АОС, которая будет представлять собой последовательность классов узлов просмотренных обучаемых, чтобы моделировать краткосрочную информацию о нем. Каждому узлу сопоставлен алфавитный символ, характеризующий его класс (алфавитное наименование класса). Понятие класса необходимо для формализации и описания, навигационных правил, и для логического вывода на них. Класс используется как компонент для представления пользовательской модели и навигационных правил.

Классы узла определяются по следующим критериям:

1) показывает ли система содержание для определенного вида пользователя;
2) предлагает ли узел пользователям объяснение, задает вопрос, предлагает выполнить тестовое задание или делает что-то еще в образовательной цели дидактического плана;
3) к каким из категорий контента АОС принадлежит узел (в случаях, когда информация содержания узла может принадлежать более чем одной категории).

История взаимодействия обучаемого с АОС представляется как последовательность классов узлов, которые он посетил и множество информации с которых он просмотрел [Шубин и др, 2012].

Навигационный метод основан на системе решения, которая управляет связями узлов и решает, какие узлы, которые основаны на навигационном правиле и могут быть связаны с текущим узлом, скрывать. Представляется целесообразным реализовать следующие четыре вида навигационных правил:

1) правила пути узла;
2) общие правила пути;
3) пользовательские правила узла;
4) общие пользовательские правила.

Навигационное правило, которое использует набор параметров обучаемого из его модели, назовем правилом пути и навигационное правило, которое использует пользовательские параметры, назовем пользовательским правилом. Навигационное правило может также быть разделено на два типа: правило узла и общее правило. Правило узла определено и применяется только для определенного узла. Общее правило – для того чтобы описать наиболее часто встречающиеся навигационные пути в гиперпространстве и часто используемые сегментации диапазона параметров обучаемого.

В навигационном правиле разработчик АОС описывает связи, которые должны быть показаны в соответствии с идентификатором узла или в соответствии с классом узла, который является целью связи. Система скрывает все связи, на которые не ссылаются в навигационном правиле.

Навигационные правила генерации стратеги обучения *S*, для адаптивной модели обучения M_2:

1. общее навигационное правило:

$$M1(P11 \wedge P1h \wedge \ldots \wedge Pm1 \wedge Pmh) \rightarrow M2(P1 \wedge \ldots \wedge Pn);$$

2. навигационное правило узла:
$M1(P11 \wedge P1h \wedge ... \wedge Pm1 \wedge Pmh) \rightarrow S\ (D1 \wedge ... \wedge Dn)$;
3. общее навигационное правило пользователя:
$e1\#S\ (M1, Me)\#e_2 \rightarrow M2(P1 \wedge ... \wedge Pn)$;
4. локальное навигационное правило пользователя:
$e_1\#S\ (M1i, Mei)\#e_2 \rightarrow S_2(D1 \wedge ... \wedge Dn)$,

где: pi – дидактический предикат множества; D – идентификатор узла, информация которого будет показана обучаемому в рамках генерации стратегии обучения; *h* – число историй, которые были задействованы в модели обучаемого M_1; m – число образов пути; n – число идентификаторов узлов, которые будут показаны обучаемому согласно модели Me ; $S(M_{1i}, M_{ei})$ – параметр степени толерантности на *i*-м шаге; e – граница степени толерантности (от 0 до 1); # – операция, которая представляется одним из следующих логических операторов: '<', '<=' или '='.

С левой стороны формулы у первом и втором правилах специфицирован образец истории пути M_1, который представляет собой результат работы пользователя в дидактическом пространстве M ИНАГС. Навигационное правило означает, что система показывает обучаемому связи, идентификатор узла или класс которых указан в его правой части, если история пути обучаемого соответствует одному из образцов историй, которые записаны в его левой части. Пользовательское навигационное правило означает, что система показывает связи, идентификатор узла или класс которых указан в его правой части, если параметр коэффициента толерантности знаний, указанный в левой части, находится в пределах заданного диапазона. Если у узла есть несколько навигационных правил, система показывает все связи, которые подтверждаются любым навигационным правилом. Это означает, что если, по крайней мере, одно правило из нескольких правил одобряет показ определенной связи, система показывает связь независимо от других навигационных правил.

Пример навигации с использованием правил пути и пользовательских правил. Классы определены следующим образом: A – узлы с вопросом (тестовым заданием); C – узлы, которые содержащие информацию для правильного ответа обучаемого; C – узлы с информацией для неправильного ответа обучаемого; D – узлы с объяснением для студентов с высоким уровнем знаний; E – узлы с объяснением для студентов с низким уровнем знаний. Навигационное правило определено для дидактического предиката P_i узла $B5$. Данное правило применяется только в этом узле. Правило $A \wedge C \wedge A \rightarrow D7$ означает, что, когда обучаемый будет «заходить» в узел $B5$ и история пути пользователя – $A \wedge C \wedge A$, система показывает связь к узлу $D7$ и скрывает связь к узлу $E8$. Поскольку класс $B2$ означает, что обучаемый ответил правильно, а классы $C1, C3$ – неправильно, история взаимодействия пользователя с ИНАГС показывает, что обучаемый отвечал на вопросы в узле $A0$ неправильно, а в узле $A4$ – правильно. Правило $A \wedge B \wedge A \rightarrow E8$ определяет связи таким образом, что если обучаемый ответил верно на вопросы в узлах $A0$ и $A4$, то система показывает только связь до узла $E8$. Таким образом, система изменяет стратегию S процесса обучения в соответствии с текущими результатами учебной деятельности пользователя.

Выводы

Предложен подход к разработке интеллектуальных обучающих систем на основе методов адаптивной гипермедиа. В качестве математического аппарата описания выбрана алгебра конечных предикатов и

предикатных операций. Предложенный подход снижает трудоемкость разработки обучающих систем за счет автоматического управления ходом учебного процесса посредством навигационных правил. Кроме того, предложены модели, реализация которых приводит к уменьшению количества ошибок в описании навигационных правил, для которых был разработан специальный формат, что делает их более простыми для восприятия. Созданная на основе предложенного подхода интеллектуальная обучающая система управляет маршрутом движения обучаемого по своему контенту с помощью технологии сокрытия связей для оптимизации учебного процесса. Также разработано инструментальное средство проверки наличия ошибок в созданных навигационных правилах с целью дальнейшего сокращения их количества. Предложенные подходы и алгоритмы могут быть использованы в любых системах адаптивной гипермедиа, которые управляют перемещением пользователя по гиперпространству.

Bibliography

[Bondarenko et al, 2008] M. Bondarenko, N. Bilous, I. Shubin. The Ukrainian e-Learning Region: In Proceedings of 10-th International LLinE Conference New Partnerships and Lifelong Learning, Helsinki, Finland, 2008.

[Shubin et al, 2011] I. Shubin, T. Gorbach, A. Scherbak, Y. Svyatkin The technique of adaptive interactive lectures for the «Multimedia Systems» course // Сборник научных трудов по материалам 13-й Междунар. конф. «Образование и виртуальность ВИРТ-2011», Харьков - Ялта, Украина - 2011.

[Святкин, 2012] Я.В. Святкин Модель обучения с применением навигационных правил генерации и динамической модификации стратегий обучения в базисе алгебры конечных предикатов// «Вестник НТУ ХПИ» Национальный технический университет «ХПИ», Харьков, Украина, 2012

[Шубин и др, 2012], І.Ю. Шубін, Я.В. Святкин Методы и модели построения интеллектуальных адаптивных гипермедиа систем/ // Восточно-европейский Журнал передовых технологий 3/11(57), Харьков, 2012

Authors' Information

Igor Shubin *– Professor of Software Department,*
Kharkov National University of Radioelectronics, Lenin ave. ,bl. 14, Kharkov, 61166, Ukraine;
e-mail: shubin@kture.kharkov.ua

Irina Kyrychenko *– Researcher; Kharkov National University of Radioelectronics, Lenin ave. ,bl. 14, Kharkiv, 61166, Ukraine;*
e-mail: ikyrychenko@mail.ru

Vladimir Gritsenko *– Student; Kharkov National University of Radioelectronics, Lenin ave. ,bl. 14, Kharkiv, 61166, Ukraine;*
e-mail: 27bebecap@gmail.com

Vladimir Chernov *– Researcher; Kharkov National University of Radioelectronics, Lenin ave. ,bl. 14, Kharkiv, 61166, Ukraine;*
e-mail: shubin@kture.kharkov.ua

TABLE OF CONTENT