METHOD OF SIMULATION OF ADAPTIVE BEHAVIOR ANT COLONY FOR THE CONTROL OF SPELLING ERRORS IN TEXTS

Tishlikov Sultonjon Abduraimovich[1]
[1]*Head of the Department of Information Technology, Doctor of Philosophy (PhD) in Technical Sciences, Associate Professor of Gulistan State University*
***e-mail**: tsa_sultonbek@bk.ru*
Bobomurodov Jasurbek Abdirasul oglu[2]
2nd year master's student, Information Systems,
*GulSU, **e-mail**: jascodersuz@gmail.com*
Egamberdiyeva Rano Suyun qizi[3]
1st year master's student, Information Systems,
GulSU, e-mail: ranoegamberdieva2@gmail.com

Abstract. *The principles and algorithms are offered for control of the Uzbek language spell on the basis of method of searching for the correct wordform by synthesis models of adaptive behavior of ant colonies. The developed search algorithms improve the performance of monitoring system in comparison with the methods of brute force dictionary of wordforms.*

Key words: *Parsing, n -gram. Coding, word-form, development, information-search, Word predictor, Tagger, Constructor, Predictor, parsing model, colony, behavior, adaptive, algorithm, Cyclic method, stochastic search.*

MATNDAGI IMLOVIY XATOLARNI NAZORAT QILISHDA CHUMOLI TO'DASINING MOSLASHUVCHAN HARAKATINI MODELLASHTIRUVCHI USULNI QO'LLASH

Annotatsiya*. Chumoli to'dasi moslashuvchan harakatini modellari sintezi bo'yicha, kerakli so'z shaklini izlovchi usul asosida o'zbek tili orfografiyasini nazorat tizimini yaratish printsip va algoritmlari taklif qilingan. Ishlab chiqilgan algoritmlar lug'atdagi so'z shakllarini qidiruvchi usullariga nisbatan nazorat ko'rsatkichlarini ancha yaxshilaydi.*

Kalit so'zlar. *Parsing, n-gram, kodlash, so'z shakli, predictor, tagger, konstruktor, bashoratchi, parsing model, koloniya, algoritm, tsiklik usul, stoxastik qidiruv.*

МЕТОД АДАПТИВНОГО МОДЕЛИРОВАНИЯ ПОВЕДЕНИЯ МУРАВЬИНОЙ КОЛОНИИ ПРИ КОНТРОЛЕ ОРФОГРАФИЧЕСКИХ ОШИБОК В ТЕКСТАХ

Аннотация.*Предложены принципы и алгоритмы системы контроля орфографии узбекского языка на основе метода поиска нужной словоформы путем синтеза моделей адаптивного поведения муравьиной колонии. Разработанные*

поисковые алгоритмы улучшают показатели системы контроля по сравнению с методами перебора словаря словоформ.

Ключевые слова. *Синтаксический анализ, n-грамм, кодирование, словоформа, разработка, поиск информации, предиктор слов, таггер, конструктор, предиктор, модель синтаксического анализа, колония, поведение, адаптивный, алгоритм, циклический метод, стохастический поиск.*

INTRODUCTION

The computer system of checking and corrections of spelling for various natural languages, constructed on the mechanism of n-gram structured modeling, is based on use of procedures which performance parsing representing, word-form coding, of adaptive search of the necessary information in the large volumes frequency dictionaries [1,2]. Moreover, use of software-tool pledged in the software of modern computers for search destructured part of word-form is not given satisfactory decisions when sizes of search attributes and specific characteristics of objects are large. In this connection the decisions of questions directed to development of methods and algorithms determining the order of automatic search of attributes by processing the information received at a level of the absolute description.

There are various approaches to construction of program modules for searching and selection of the necessary attributes, specific characteristics and for word-form description in system of spelling checking and correction. Among them it is possible to allocate heuristic, information, statistical and probability approaches to processing the dictionaries of word-forms. In practice the decision of information-search tasks by full sorting methods are most spread, and they are highly iterative, connected to necessity of combinational search of acceptable integration of attributes, characterized by large expenses of machine time and appear poorly suitable at the decision of tasks with large dimension.

One of effective methods for decision of search tasks with combinational nature is the stochastic method, which allows to carry out search faster than many traditional sorting methods and does not require sorting of all possible combinations [2]. The probability-stochastic search of object can be presented as a task of planning based on application of tree-decisions, i.e. on methods of artificial intelligence with properties of self-adapting and self-organizing.

Let's note, that the adaptation on the basis of self-training and self-organizing dominates in the existence and evolution of living organisms. In this connection, using the methods of intellectual analysis and processing of information at dynamic systems on the basis of modeling the biological systems are represent the large theoretical and practical interest, for example, modeling of behavior of ant colony, capable quickly find the shortest way from the anthill to food source. The colony is e represented as sensible multiagent system ensuring achievement of joint purposes on the basis of low-level interaction. The algorithm of modeling of ant colony behavior makes a basis of self-organizing and specialty of models is the availability of indirect exchange, which is used in algorithms. The indirect exchange represents the interaction, separated in time, at which one individual agent changes some area of an environment, and others use this information later, when they catch in area [3].

In section below we state technique of designing the algorithm for objects search and for selection of composite components of word-form at the control and correction of spelling mistakes on the basis of the parsing mechanism and adaptive ant system.

PARSING MODELING OF WORD STRUCTURE

The mechanism of applying the n-gram structured model of natural language includes procedures of parsing coding and search the sequence of controllable words. The principle of parsing coding is the following.

Let W be the sentence consisted from n words. Begin of sentence we designate as $< s >$ and end of sentence as $< / s >$, so we have $w_0 =< s >$ and $w_{n+1} =< / s >$. If $W_k = w_0 ... w_k$ is number of k-prefixes of word in the sentence, then $W_k T_k$ is k-prefix of word-parsing. For coding a sequence of words we have to construct the tree of word – parsing [4].

Note, that k-prefix of word-parsing contains only those binary sub-tree, which diapazones are completely included in k-prefixes of word excepting of $w_0 =< s >$. The separate words with their positional attributes (POS-attribute) can be accepted as root trees.

In fig. 1 we illustrate complete parsing of some word. The circuit determines binary parsing $(< s > SB)(w_1, t_1)...(w_n t_n)(< / s >, SE)$, where sequence SB / SE is the

distinctive POS-attribute for $< s > / < / s >$, according to restrictions, that $(< / s >, TOP)$ is single legal heading; $(w_1, t_1)...(w_n, t_n)(< / s >, SE)$ forms an element with heading $(< / s >, TOP')$.Parsings are defined, when $(< / s >, TOP')$ is heading of any element, which dominates above $< / s >$, but not $< s >$.

In fig. 2 we submitted the circuit of coding system modules interaction for constructing the algorithm of elements recognition on the basis of parsing tree.

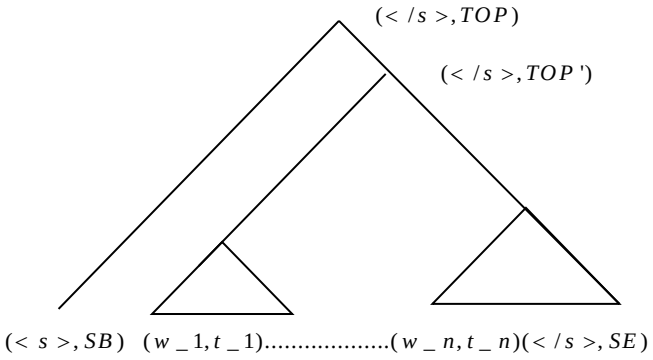


Fig. 1. Complete parsing

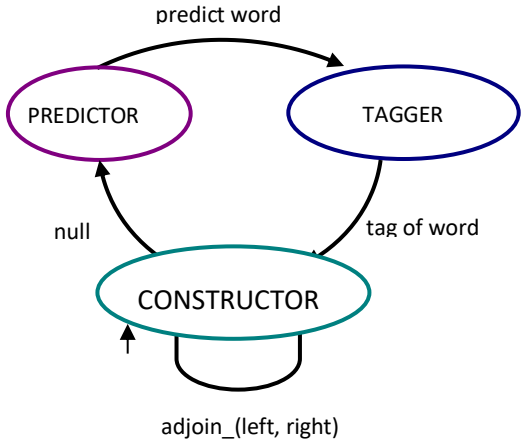


Fig. 2. Interaction of modules at parsing coding system

The system of coding consists from three modules:

1) "Word predictor" predicts the following word w_{k+1} given by k -prefix of word-parsing, then passes control to "Tagger";

2) "Tagger" predicts the POS-attribute t_{k+1} of the next word, given k -prefix of word-parsing and the newly predicted word w_{k+1}. Module passes control to "Constructor";

3) "Constructor" grows the already existing binary branching structure by repeatedly generating transitions until module passes control to "Predictor" by taking a null transition.

The follow result of research is devoted to getting the estimation of information interchange probabilities between modules of parsing model.

Probability estimation for search of the word on the parsing model. Let's designate the probability of search and recognition of word sequence W in parsing model as $P(W, T)$, where T is the tree of complete parsing. The probability

model should be capable distinguish desirable and less desirable parsings. To receive the correct assignment of probability $P(W,T)$ it is necessary to define appropriate conditional probabilities for each transition.

The probability $P(W,T)$ of the word sequence W and complete parsing T is calculated as follows:

$$P(W,T)=\prod_{k=1}^{n+1}[P(w_k|W_{k-1}T_{k-1})\cdot P(t_k|W_{k-1}T_{k-1},w_k)\cdot P(T_{k-1}^k|W_{k-1}T_{k-1},w_k,t_k)].$$

Here $P(T_{k-1}^k|W_{k-1}T_{k-1},w_k,t_k)=\prod_{i=1}^{N_k}P(p_i^k|W_{k-1}T_{k-1},w_k,t_k,p_1^k\ldots p_{i-1}^k)$, $W_{k-1}T_{k-1}-(k-1)$ is prefix of word-parsing; w_k is word predicted by "Word-Predictor"; t_k is the tag assigned to w_k by "Tagger"; T_{k-1}^k is the incremental parsing structure that generates $T_k=T_{k-1}\|T_{k-1}^k$ when parsing structure is built on top T_{k-1} and newly predicted word w_k; the notation $\|$ stands for concatenation; N_{k-1} is the number of operations "Constructor" executes at position k of the input string before passing control to "Word-Predictor" (N_k-th operation at position k is the null transition), N_k is a function of T; p_i^k denotes the i-th "Constructor" action carried out at position k in the word string as follow:

$p_i^k \in \{(adjoin-left, NTag), (adjoin-right, NTag), (uniray, NTag)\}$,

$1\le i<N_k$, $p_i^k=null$, $i=N_k$.

Note that each $(W_{k-1}T_{k-1},w_k,t_k,p_1^k\ldots p_{i-1}^k)$, $i=1,\ldots,N_k$ defines valid k-prefix of word-parsing W_kT_k at position k in the sentence.

Now let's state results of developing the algorithm of object search for control and correction the spelling mistakes.

Formalization of search model. We formulated the problem of object search as follow. There is a set of modules $M=\{m_i|\ i=1,2,\ldots,n\}$ which are the frequency dictionaries and dictionaries of word-forms. Each module is characterized by three

elements $< S_i, l_i, t_i >$, where S_i is area of the module, the parameters l_i and t_i set the bottom and top border of value h_i / w_i, i.e.

$$l_i \le h_i / w_i \le t_i \qquad (1)$$

where h_i is the height of module, w_i is the width of module.

For set of modules M we make the scheme-plan, similar to the decisions of planning tasks. The scheme-plan represents rectangular R cut by vertical and horizontal lines on set of areas r_i in each of which the module m_i accordingly is located [5].

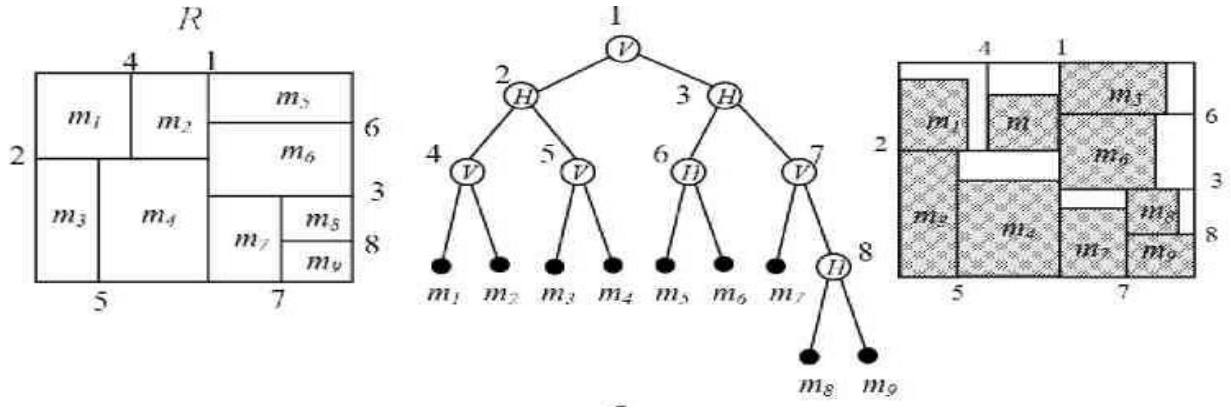
On the tree the tops appropriate to sections are marked by figures, where V is vertical section and H is horizontal section. The letters mark tops appropriate to areas. Each area r_i, intended for accommodation of module m_i, has the sizes x_i and y_i.

Having observance subject of condition (1) the sizes of area should also correspond to conditions $S_i \le x_i \cdot y_i$, $h_i \le y_i$, $w_i \le x_i$

(2)

The aim of optimization is the minimization of total plan area at observance of conditions (1) and (2).

In fig. 3-a) we submitted the plan and in fig. 3-b),c) we illustrated appropriate to it tree $D = \{d_j \mid j = 1,2,..., 2n-1\}$ as the leaf of which are the tops appropriate to blocks, and the internal tops correspond to sections: V is vertical, H is horizontal.



a) plan of search section b) tree of search section c) plan with located modules

Fig . 3. Scheme-plan for object search

Search of object on the basis of convolution method. Let's present structure of sections tree as a sequence of binary sections. For internal tops of a tree appropriate to sections we marked the type of section H or V, the leaf of tree is indicated by numbers of modules, for modules with the fixed sizes is underlined their orientation.

On the basis of this information the construction of object search plan is carried out by consecutive binary convolution of areas on sections tree, beginning from leaf of tree. To each internal top of sections tree is correspond area got as a result of binary convolution of sub-tree which has the root in internal top.

We assumed that section with number i cut top d_i (area u_i). In the beginning of convolution to each top d_i being a leaf of sections tree, we put in conformity the area u_i with the sizes $x_i = h_i$, $y_i = w_i$, equal to the sizes of module m_i. Let tops d_i and d_j are affiliated tops of d_k and let for areas u_i and u_j, appropriate d_i and d_j, the bottom borders of their sizes (x_i, y_i), (x_j, y_j) are determined.

The offered way of binary convolution represents merge of areas u_i and u_j, formation of area u_k, definition of the sizes for u_k and new sizes for u_i and u_j.

Further we designate through $\max(x_1, x_2)$ the maximal value of x_1 and x_2. At merge on horizontal the values of $y_k = \max(y_i, y_j)$; $x_k = x_i + x_j$; y_i and y_j have the

size equal $\max(y_i, y_j)$. At merge on vertical the values of $y_k = y_i + y_j$; $x_k = \max(x_i, x_j)$; x_i and x_j have the size equal $\max(x_i, x_j)$.

For example from fig. 3-a) at the fixed sizes of modules $(m_1 - m_9)$ the plan with the placed in areas modules after metrization looks like to submitted scheme on fig. 3-c) . It is necessary to differ the sizes (h_i, w_i) of the module m_i from the sizes (x_i, y_i) of area u_i, in which the module (block) m_i is placed. The sizes of areas and describing rectangular of the plan are defined by consecutive convolution of areas on section tree, proceeding from the sizes of modules placing in these areas.

The description of structure of object search tree. Let's enter the alphabet $A = \{M, R\}$. Structure of the search tree for word-form in Uzbek language is set on the basis of alphabet A as expression for a binary tree, where the set $M = \{mi \mid i = 1,2,..., n_x\}$ corresponds to leaf of section tree (areas), and the set $R = \{H, V\}$ corresponds to sections. Uzbek word-form on the tree submitted in fig. 3-b) has the view:

$$D = m_1 m_2 V\ m_3 m_4 V\ H\ m_5 m_6 H\ m_7 m_8 m_9 H\ V\ H\ V\ .$$

The process of restoration of Uzbek word-forms tree is simple enough. The word is looked through consistently from left to right and the appropriate to sections letters H or V are found. Each such section unites two nearest subgraph formed on the previous steps and located in Uzbek word to the left of letter H or V. Process of convolution with the help of brackets results as:

$$D = ((m_1 m_2 V)(m_3 m_4 V) H)(m_5 m_6 H)(m_7 (m_8 m_9 H) V) H) V\ .$$

We noted the basic properties of Uzbek word description, which performance is necessary for corresponding the section tree to record.The number of elements belonging M is designated as n_M, and number of elements appropriate to sections as M

1. For each module m_i two ways (two orientations) of locating in area u_i are possible. To first orientation of module m_i designation m_i^1 is corresponds, and to second orientation m_i^2 is corresponds.

2. Structure of expression includes all elements of set $M = \{m_i \mid i = 1,2,..., n_M\}$ from one of labels m_i^1 or m_i^2.

For sections tree the equation $n_M = n_R + 1$ is always carried out.

If to drow on the right of letter H or V the cut, then to the left of cut the number of elements belonging to set M will be more than number of elements appropriate to sections, at least, on 1.

At viewing expression from left to right, the first element appropriate to sections in expression can appear only after two elements belonging to set M.

Expression D, constructed on the basis of the alphabet we named as legitimate if it satisfies to above conditions. Thus, the legitimate expression D is symbolical representation of the decision of object search task. Note, that the various decisions is got by combination the relative position of alphabet $A = \{M, R\}$.

The decision of object search task by representation of solutions space as set of legitimate expressions D is as follow. The search of decision is reduced to search such legitimate expression D, which optimizes a parameter (criterion) of quality and is carried out on the complete graph $G = (X, U)$, where $X = X_1 \cup X_2 \cup X_3 \cup X_4$. The tops of set $X_1 = \{x_{1i} \mid i = 1,2,..., n_M\}$ correspond to modules m_i^1 placed in the first orientation. The tops of set $X_2 = \{x_{2i} \mid i = 1,2,..., n_M\}$ correspond to modules m_i^2 placed in the second orientation. The tops of set $X_3 = \{x_{3i} \mid i = 1,2,..., n_R\}$ correspond to horizontal sections H. The tops of set $X_4 = \{x_{4i} \mid i = 1,2,..., n_R\}$ correspond to vertical sections V.

Now let's state the decisions of a task on the basis of modeling the behavior of ant colony.

Organization of object search on the basis of adaptive ant colony behavior. Let's assumed that the search of object is carried out by collective of ants $Z = \{z_k \mid k = 1,2,..., l\}$. On each iteration of ant-search algorithm each ant z_k builds the concrete solution of search. The solution is the route in graph $G = (X, U)$, which

includes n_M tops belonged to sets X_1 or X_2, and n_R tops belonged to sets X_3 or X_4. In this case constructed route is represented as legitimated expression D.

For uniform distribution of ants and the creations of equal starting conditions as initial tops at routes, formed by ants, are used tops of sets X_1 and X_2 with total quantity $2 \cdot n_M$. In other words, the number of solutions formed by ants on each iteration is equal $2 \cdot n_M$. It is connected to distribution of ferment on graph G edges. At first it is considered, that on all graph G edges the identical (small) quantity Q / v of ferment is postponed, where $v = |U|$. Parameter Q is set a priory. Process of search solutions is iterative. Each iteration l includes three stages. At the first stage the ant finds the solution, at the second stage the ant postpones the ferment, at the third stage the evaporations of ferment is carried out. Further we stated the cyclic method for stochastic search of object by ant-search algorithms.

CYCLIC METHOD OF STOCHASTIC SEARCH

We considered a case, when ferment is postponed by the agent on edges after complete formation of solution. At the first stage of each iteration everyone k-th ant forms an own route D_k. Process of constructing the route D_k is step-by-step. On each step t the agent applies probabilistic rule of choice the following top to inclusion it in formed route $D_k(t)$. For this the set of tops $X_k(t) \in X$ is formed so, that each of tops $x_i \in X_k(t)$ can be added in the formed route $D_k(t)$.

Now let $e_k(t)$ be the last top of route. The agent looks through all tops $x_i \in X_k(t)$. For each top $x_i \in X_k(t)$ the parameters f_{ik} is calculated and they are total level of ferment on graph G edge, connecting x_i to top $e_k(t)$. The probability P_{ik} of inclusion of top $x_i \in X_k(t)$ in a formed route $D_k(t)$ is defined by the following parity

$$P_{ik} = f_{ik} / \sum_i f_{ik} . \qquad (3)$$

The agent chooses with probability P_{ik} one from tops, which is included in the route $D_k(t)$.

At the second stage of iteration, each ant postpones ferment on edges of the constructed route. The quantity of ferment $\Delta\tau_k(l)$, postponed by ant z_k on each edge of the constructed route, is defined as

$$\Delta\tau_k(l) = Q / F_k(l), \quad (4)$$

where l is the number of iteration, Q_i is total quantity of ferment postponed by ant on edges of the route D_k, $F_k(l)$ is criterion function for the decision received by ant z_k on l-th of iteration. Than $F_k(l)$ is less, then more ferment is postponed on edges of the constructed route and and, hence, the probability of choice of these edges at construction of routes on next iteration is more.

At the third stage the full evaporation of ferment on edges of complete graph G is made according to the formula

$$f_{ik} = f_{ik}(1-\rho), \quad (5)$$

where ρ is the coefficient of updating.

After performance of all actions on iteration the agent with the best decision is finding and it is remembered. Then the transition to the following iteration is carried out.

Let's note, that the time difficulty of developed search algorithm on the offered method depends on number of iterations l, quantities of graph tops n, numbers of search images (agents - ants) m and is defined(determined) as $O(l \cdot n^2 \cdot m)$.

CONCLUSION

The carried out experimental researches of developed system for control and correction of spelling mistakes at large volume of context in the Uzbek language have allowed to create space of decisions. The search process is organized on the basis of modeling of adaptive ant colony behavior. The comparison with known algorithms was shown, that for decisions, received with the help of ant-search algorithm, the operating time is smaller and at the same time the value of criterion function is better (less) on 6 % in average. On average, the start of program provides finding of the decision distinguished from optimum less than on 1 %.

REFERENCES

[1]. Jumanov I.I., Akhatov A.R., Kurbanov M.M., Karshiev Z.A. Use of N-gram statistics for checking of the texts transfer quality in intellectual information systems // The Fifth World Conference on Intelligent Systems for Industrial Automation, 2008. – Tashkent, ISBN 3-933609-27-5, b-Quadrat Verlag-86916 Kaufering, – 153-160 p.
[2]. Jumanov I.I., Tishlikov S.A. The control of the text information transfer and processing reliability on the basis of technology of parallel computing // Proceedings international training-seminars on mathematics, SamSU and Malaysian Mathematical sciences society, Samarkand, 2011, p.p. 215-217.
[3]. Ferreira C. Gene Expression Programming: A New Adaptive Algorithm for Solving Problems // Complex Systems, 2001, Vol. 13, issue 2: 87-129.
[4]. CiprianChelba, Frederick Jelinek. Structured language modeling // Computer Speechand Language (2000) 14, 283–332.
[5]. Lakhmi C. Jain, Martin N.M. Fusion of Neural Networks, Fuzzy Systems and Genetic Algorithms: Industrial Applications. — CRC Press LLC, 1998. – 368 p.
[6]. Jumanov I.I., Karshiev Kh.B., Tishlikov S.A. Examination of the Efficiency of Algorithms for Increasing the Reliability of Information on Criteria of Harness and the Cost of Processing Electronic Documents// International Journal of Innovative Technology and Exploring Engineering(TM) Volume-9 Issue-1, November 10, 2019. p.p. 4133- 4139.(№3;Scopus;IF=0.6).
[7]. Akhatov.A.R, Tishlikov S. A., Isroilov N.A. Control of authenticity of iformation transfer and processing on the basis of mechanisms for adjustment of transients identification models parameters // In proceedings of the Ninth World Conference on Intelligent Systems for Industrial Automation, 25-27 October, 2016. – Tashkent, Uzbekistan, 2016. – P.211-215.
[8]. Jumanov I.I., Tishlikov S. A. Control of integrity and authenticity of electronic documents on the basis of genetic principles of tests formation and generation // In proceedings of the Eight World Conference on Intelligent Systems for Industrial Automation, 25-27 November, 2014. – Tashkent, Uzbekistan, 2014. – P.242-246
[9].Jumanov I.I., Tishlikov S. A. The system of transmitted text information control on the basis of identification of genetic algorithms // In proceedings of the Seventh World Conference on Intelligent Systems for Industrial Automation, 25-27 November, 2012. – Tashkent, Uzbekistan, 2012. – P.196-200
[10]. Jumanov I.I., Tishlikov S. A. Increase of efficiency of word-forms search and processing for the control and correction of spelling mistakes in the electronic texts // The International Scientific Conference «Modern problems of applied mathematics and information technologies – Al-Khorezmiy 2012», Volume № 2, 19-22 December, National University of Uzbekistan. - Tashkent, 2012. – p.p. 93-96