ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 9 , September 2019

Increasing information validity using natural redundancy and n-grams of natural language

Jumanov Isroil Ibragimovich, Karshiev Khusan Berkinbaevich

Doctor of Technical Sciences, Professor, Department of Information Technologies, Samarkand State University, Samarkand, Uzbekistan
Basic Doctoral Student, Department of Information Technologies, Samarkand State University, Samarkand, Uzbekistan

ABSTRACT: The problem is formulated and the methodological foundations for creating a technology for increasing the reliability of information in full-text documents of electronic document management systems of institutions using statistical, natural, semantic information redundancy are developed. Optimal information processing tools based on n-grams of structured natural language models have been designed. Methods have been developed for calculating the conditional entropy of document elements from a large collection, determining the amount of natural redundancy required to ensure the accuracy of information. The regularities of the distribution of the frequency characteristics of n-grams are investigated, and also dictionaries of n-grams and mechanisms for using typical search, recognition, and clustering tools are constructed. Mathematical expressions of the probability of detection, non-detection of errors and analysis of the reliability of information are obtained. A software package for detecting and correcting k– multiple information errors was developed, which was implemented and tested in the framework of the Sphinx-4 framework.

KEY WORDS: electronic document management systems, information reliability, information redundancy, probability of non-detection of errors, recognition, clustering, n - gram model, software package.

I. INTRODUCTION

Creating electronic document management systems (EDMS) has been developing for many years in many countries and certain successes have been achieved [1-3].

Technologies of EDMS are defined by a wide range of applications such as e-mail readers, spelling control editors for natural languages, machine translation tools, speech and text synthesizers and analyzers, office documents processing, programs that can read by voice, play text files, etc. [4].

In the noted area of research, there are technologies from Microsoft, Lucent, Lernout & Hauspie, Unisys, Elan, and others. Unisys Corporation software packages are also used that are aimed at recognizing and “understanding” human speech, as well as conducting a full-fledged dialogue with a computer [5,6]. In addition, there are standard software for EDMS, aimed at solving, first of all, the tasks of accounting, monitoring the implementation of organizational and administrative documents (OAD), analysis and search, the formation of databases (DB) and knowledge bases (KB) [7].

The key among them is the problem of increasing the reliability of information in electronic documents (ED). Information is distorted when documents are entered through the fault of the human operator, due to errors in scanning and recognition devices, the influence of interference in communication channels, and also due to the presence of spelling errors made by users of the system.

The use of erroneous information, in turn, reduces the effectiveness of the system and leads to the loss of value, accuracy, completeness, its temporality of the document and information almost completely. Common methods for ensuring the reliability of information to the required level are modern data transfer protocols that use corrective codes that allow you to detect and correct errors by interrogating [8]. Traditional software methods that are used in EMDS are far from perfect and increasing the reliability of information to the required level is achieved with significant costs. In addition, they do not justify themselves for reasons of high cost, narrow specialization of tools [8-10].

A prerequisite for solving this problem is the improvement of traditional and the creation of new tools for monitoring and correcting errors in texts, performed by significantly less labor and the cost of processing information. A promising direction for increasing the reliability of information is the development and implementation of algorithms based on the typical functions of search, recognition, classification, generation, translation of texts from one language to another, and error control and correction programs of various multiplicity based on the use of natural information redundancy[11].

There is a wide range of works in this area, included in the range of the most developed studies, which use spelling errors control methods based on the implementation of graphematical, linguistic, morphological, n-gram structured NL models [12].

Features that are the attraction of a large wordform dictionary, the application of mechanisms for extracting specific knowledge in documents and the detection and correction of errors of various multiplicities [10-12]. The regularities of the distribution of n-gram errors based on a large body of materials used in the activities of public services institutions were studied and modelled, and algorithms for fixing distorted elements (letters, words) in the text of documents, search, recognition, clustering and structuring tools, as well as the use of frequency characteristics of information, dictionaries of n-grams of small volume. Error Detection and Correction Approach k-multiplicity based on n-grams is based on the experiments of C. Shannon, in which the results are presented on the basis of the use of diagrams. Due to the presence of natural information redundancy, the most likely letter follows from the letter x_i, and such an event is predicted from the set $x_{i-1}, x_{i-2}, \ldots, x_{i-n}$. In [4], a technique was developed for calculating the conditional entropy for the set of (k + 1) matrix elements (k-grams). Moreover, for small k the following condition must be fulfilled:

$$H(x_{ij} / x_{ij}^{(1)}, x_{ij}^{(k)}) = H(x_{ij}, \overline{x_{ij}^{(1)}, x_{ij}^{(k)}}) - H(x_{ij}^{(1)}, x_{ij}^{(k)});$$

$$H(\|x_{ij}\|) = H(x_{ij}, \overline{x_{ij}^{(1)}, x_{ij}^{(k)}}) + \sum_{r=k+1}^{ns-1} \min H(x_{ij}^{(r)} / x_{ij}^{(1)}, x_{ij}^{(k)}). \qquad (1)$$

The above relations give fairly close values of the estimation of conditional entropy. Estimates of the conditional entropy of k-grams are found as

$$\tilde{H}(\overline{x_{ij}^{(1)}, x_{ij}^{(k)}}) = -\sum_{\xi_{ij}^{(1)} \xi_{ij}^{(k)}} \tilde{P}(\overline{\xi_{ij}^{(1)}, \xi_{ij}^{(k)}}) \log \tilde{P}(\overline{\xi_{ij}^{(1)}, \xi_{ij}^{(k)}}), \qquad (2)$$

where $\overline{\xi_{ij}^{(1)}, \xi_{ij}^{(k)}} \in \overline{x_{ij}^{(1)}, x_{ij}^{(r)}}$ is a concrete implementation of k – grams;

$\tilde{P}(\overline{\xi_{ij}^{(1)}, \xi_{ij}^{(k)}})$ -is an estimate of the probability of occurrence of k – grams.

The conditional entropies of k - grams are calculated on the basis of a collection of 100 documents and the characteristics of monograms, diagrams, and trigrams of information are obtained. The amount of redundancy required to increase the reliability of information is determined, the value of which ranges at $0{,}5 \div 0{,}7$.

II. EFFICIENCY EVALUATION CRITERIA FOR METHODS OF IMPROVING INFORMATION VALIDITY

The possible number of distortions in the sequence of information encoded by a control code (CC) depends on the number of code words, and their estimates are presented as a function

$$f_{KK_i} = \sum_{m=1}^{n} f_{KK_i}^{m}, \qquad (3)$$

where $f_{KK_i}^{m}$ is the number of possible combinations of code combinations caused by k-fold errors.

In expression (3), the value $f_{KK_i}^m$ acts as an indicator that characterizes the ability of the control code to detect erroneous positions of letters or characters in the text. However, the f_{KK_i} function only takes into account the total number of information errors without dividing them into single, double, and k-fold errors. It is important to note that the function f_{KK_i} does not take into account the specific gravity of each type of error and the frequency of their occurrence.

An approach is proposed by which the effectiveness of methods to increase the reliability of information is estimated by the generalized value F_{KK_i} , which is proposed as a criterion for the probability of undetected errors

$$F_{KK_i} = \sum_{m=1}^{n} F_{KK_i}^m \ . \tag{4}$$

Here $F_{KK_i}^m$ - the estimates of undetected erroneous elements of the document caused by m - multiple errors.

The general view of the probability of undetected errors is defined as

$$F_{KK_i}^m = \sum_{m=1}^{n} f_{KK_i}^m P^m \ , \tag{5}$$

where P^m is the a priori probability of m-fold errors in the information of ED.

The probability of a single error is denoted by P^1, which is assumed to be an independent event. Then the probability of two multiple errors is equal to the product of the probabilities of single errors, as $P^2 = \left(P^1\right)^2$. The probability of triple errors is estimated as $P^3 = \left(P^1\right)^3$, and the probability of *m*-fold errors in the form $P^m = \left(P^1\right)^m$. Based on the above manner, expression (5) is converted to

$$F_{KK_i} = \sum_{m=1}^{n} f_{KK_i}^m \left(P^1\right)^m \ . \tag{6}$$

It is legitimate to assume that the probability of a single error in the document information is $\approx 2 \cdot 10^{-4}$. Then, the reliability of one document concept, including *t* information elements, is estimated by probability $\approx 2 \cdot 10^{-4} t$. Based on these principles, we rewrite formula (6) in the form

$$\begin{aligned} F_{KK_i} &= 2 \cdot 10^{-4} t^1 f_{KK_i}^1 + 2^2 \cdot 10^{-8} t^2 f_{KK_i}^2 + 2^3 \cdot 10^{-12} t^3 f_{KK_i}^3 + \\ &+ 2^4 \cdot 10^{-16} t^4 f_{KK_i}^4 + \ldots + 2^m \cdot 10^{-4m} t^m f_{KK_i}^m \end{aligned} \tag{7}$$

$$F_{KK_i} = \sum_{m=1}^{n} \left(2 \cdot 10^{-4}\right)^m t^m f_{KK_i}^m \ . \tag{8}$$

An analysis of expression (7) shows that determining and using in the further calculations the number of erroneous positions caused by more than three-fold errors does not make sense, since the probability of their occurrence is so small that it practically reduces to zero all other terms starting from the fourth. As a result, the reliability of the document element is evaluated as

$$F_{\kappa\kappa_l} = \sum_{m=1}^{3} (2 \cdot 10^{-4})^m t^m f_{kk_l}^m . \tag{9}$$

Further, substituting the values of $f_{\kappa\kappa_i}^{m}$ into formula (9), we obtain an estimate of the reliability of the information in the concept of the document. Moreover, the value of the $f_{\kappa\kappa_i}^{m}$ indicator can be determined both analytically and on the basis of experimental studies. The smaller the value of the $F_{\kappa\kappa_l}$ value, the more effective is the method of increasing the reliability of information. The implementation of the stated theoretical provisions is aimed at using the natural redundancy of information, which leads to the development of mechanisms for identifying distorted elements (letters, words) in the text of the ED and the use of typical search tools, recognition, clustering, structuring, as well as determining the frequency characteristics, distribution laws of n-grams in the structure context and mechanism for controlling and correcting spelling errors. A technique has been developed for calculating the amount of natural redundancy necessary for an algorithm to increase the reliability of information.

In solving this problem, we use the obtained estimates of the conditional entropy of elements of the document S_i, which for application is defined by the average entropy alphabetically, provided that the previous element $\mathrm{S}_{\mathrm{k}}^{/}$ is known

$$H(S_i/S_k^{/}) = -\sum_{k=1}^{2^m} P(S_k^{/}) \sum_{i=1}^{2^m} P(S_i/S_k^{/}) \log P(S_i/S_k^{/}) . \tag{10}$$

To calculate the entropy (10), we also used the results of experimental studies to determine monograms, diagrams, trigrams, and other multidimensional probabilities of the form $P(S_i/S_k^{/})$.

An upper bound is obtained for estimating conditional entropy, which takes into account both the uneven distribution and the correlation between the document elements. An estimate of the average amount of information contained in one element of a document is given in the form

$$J(S_i/S_k^{/}) = H(S_i) - H(S_i/S_k^{/}) = -\sum_{i=1}^{2^m} P(S_i) \log P(S_i) + \sum_{k=1}^{2^m} P(S_k^{/}) \sum_{i=1}^{2^m} P(S_i/S_k^{/}) \log P(S_i/S_k^{/}) \tag{11}$$

Estimates of the amount of information suitable for any recoding of document elements are obtained in the form

$$\begin{aligned} J(S_i/S_k^{/}) &= m + \frac{Q(S)}{1-P_0(S)} \log Q(S) - \frac{Q(S)}{1-P_0(S)} \log[1-P_0(S)] + \frac{P_H(S)}{1-P_0(S)} \times \\ &\times \log P_H(S) - \frac{P_H(S)}{1-P_0(S)} \log[1-P_0(S)] - \frac{P_H(S)}{1-P_0(S)} \log(2^m-1) = m - \frac{P_H(S)}{1-P_0(S)} \times \\ &\times \log(2^m-1) - \log[1-P_0(S)] + \frac{Q(S)}{1-P_0(S)} \log Q(S) + \frac{P_H(S)}{1-P_0(S)} \log P_H(S) \end{aligned} \tag{12}$$

We note the characteristic moments of the results:

– for probability $P_H(S) = 0$ we have $J(S_i/S_k^{/}) = m$, since $Q(S) = 1 - P_0(S)$;

– under another boundary condition, we have $Q(S) = 0$, since $P_H(S) = 1 - P_0(S)$ and the amount of information $J(S_i/S_k^{/}) = m - \log_2(2^m - 1)$;

– expression $J(S_i/S_k^{/})$ defines the minimum of additional information that is necessary to eliminate the loss due to errors;

– for the code, when *n* = *7*, *k* = *3*, the reliability of the information is estimated with probability $Q(S) = (1-P)^m$, $m = 10$.

The probability of detecting errors in the information is in the form

$$P_0(S) = (1-P)^7 + 7P^4(1-P)^3. \tag{13}$$

The probability of undetected errors in information is estimated as

$$P_H(S) = (2^m - 1)P^4(1-P)^3. \tag{14}$$

The amount of information required to increase its reliability is determined as:

$$\begin{aligned} J(S_i / S_k^{/}) = m\frac{(2^m-1)P^4(1-P)^3}{[1-(1-P)^7+7P^4(1-P)^3}\log(2^m-1) - \log[1-(1-P)^7+7P^4(1-P)^3+ \\ +\frac{(1-P)^m}{[1-(1-P)^7+7P^4(1-P)^3}\log(1-P)^m + \log(2^m-1) - \log[1-(1-P)^7+7P^4(1-P)^3+ \\ +\frac{(1-P)^m}{[1-(1-P)^7+7P^4(1-P)^3}\log(1-P)^m + \frac{(2^m-1)R^4(1-P)^m}{[1-(1-P)^7+7P^4(1-P)^3}\log(2^m-1)P^4(1-P)^3. \end{aligned} \tag{15}$$

Given that $\mathrm{P} << 1$, this expression can be rewritten in the form:

$$J(S_0 / S_k^{/}) = m - (2^m-1)\log(2^m-1) - \log 7P^4(1-P) + (2^m-1)\log[(2^m-1)P^4]. \tag{16}$$

Substituting the maximum and conditional entropy estimates into formula (16), we determine the redundancy of information in the form

$$R(S_i) = 1 - \frac{1}{(1-P)}\log\frac{(2^m-1)}{nP}. \tag{17}$$

III. INFORMATION ERROR DETECTION MECHANISMS

The first principle of applying the developed mechanism for detecting distorted elements of a document is the use of a dictionary of word forms of the reference NL, typical search tools, recognition, clustering of words and comparison of the control word with the word form in the dictionary. To increase the reliability of information, it is assumed that the images - letters or words have already been received. It is established that this image is in the dictionary. If its presence is not established, then it is considered that such an image is absent in the dictionary of word forms. Next, the words closest to him are determined. The algorithm is associated with the determination of the rational volume of the required frequency dictionaries n - grams. As a result of this, the search time for a word is significantly reduced compared to a search algorithm with enumeration of all options in an excessively large volume of a word form dictionary, which slows down the control and correction of errors in document information. It is determined that in any EH the first thousands of the most frequent words covers from 70 to 90% of the entire studied context.

To modify the algorithm and speed up the search, recognition, and classification processes, a mechanism has been implemented for ordering words in frequency dictionaries according to the frequency of the words encountered. The most frequent words are singled out in a separate class and beyond. Checking the word for presence in this list even at the initial stage of the algorithm significantly accelerates the processes.

The second principle of improving the algorithm is based on the use of a five-level model of morphological analysis with an n - gram structured by complementing with a mechanism for using frequency dictionaries of word forms based on stochastic modeling by the Markov chain. The implementation of the algorithm is also associated with the application of a mechanism for dividing a fixed corpus of text into a predetermined number of pseudo-grammatical classes according to the conditional diagram probability of succession of words one after another. The mechanism of using three, four, and so on up to n-gram sequences is applied. It is determined that the use of n-gram sequences of a higher rank in the context of the material requires a greater amount of computation in the search, recognition and classification. Diagrams, trigrams in some cases become more preferable than high-ranking models, the use of which is small, which added to the increase in the reliability of information compared to trigrams.

To separate the context of materials into disjoint classes and clustering words, a mechanism is proposed that is aimed at one-sided viewing of a line of text on the left or on the right. In the case of applying a two-sided model, a line of text is viewed alternately on the left and on the right [12]. It is revealed that the word clustering mechanism based on the one-sided model allows one to ensure the selection of words and their breakdown into classes much faster without noticeable losses.

IV. ALGORITHM FOR INCREASING INFORMATION VALIDITY WITH CLUSTERING MECHANISM BASED ON N-GRAM

A principle is proposed by which the corpus of words is reduced to some extent and each of the N_v words is mapped into N_c classes, $N_c < N_v$.

Mapping a word into classes is represented as a model

$$w \to C = C(w)\,, \tag{18}$$

where w - a word can belong to only one class *C*.

Clustering optimization is evaluated by the criterion of greatest likelihood, at the beginning, based on diagrams, trigrams, etc.

The probability of clustering by a one-sided word search model is represented as

$$P(w_i) = P(w_i \setminus C(w_{i-n+1}),\ldots,C(w_{i-1}))\,. \tag{19}$$

According to model (19), the current word is processed depending on previous words that are mapped to word classes. The computational scheme of a one-sided model is used as a classifying function for all word positions with little time. The implemented clustering mechanism is aimed at performing the following procedures: each word is placed in accessible classes. The configuration is chosen according to which the value of the conditional probability of diagrams in the set of classes increases. Whenever a word moves to a new class, then the counters of other classes are not affected. Only those counters that suggest moving w_i from class C_i to class C_k, or only those counter diagrams where this word w_i belongs, are updated. The counter update equation is written as

$$\forall w : N(c_j, w) = N(c_j, w) - N(w_i, w)\,;$$

$$\forall w : N(c_k, w) = N(c_k, w) - N(w_i, w)\,;$$

$$N(c_j) = N(c_j) - N(w_i)\,;$$

$$N(c_k) = N(c_k) + N(w_i)\,,$$

where $N(c_k, w) = \sum_{\forall i: w_i \in c_k} N(w_i, w)$.

There is virtually no explicit search procedure used when updating the counters. This is because only the counters of those words that follow w_i are affected. To calculate the likelihood function, it is assumed that the controlled word w_i belongs only to the class of words C_i and the conditional probability of the word belonging to it linearly depends on the number of classes. The procedures of the clustering mechanism are initialized by displaying the most frequent words $(N_c - 1)$ with dictionary word forms, as a result of which each word belongs to its own unique class. And all the remaining words are placed in the N_c -the class.The study was conducted on the basic materials given by the sequence of the following classes: 20, 50, 100, 200, 500 plus in each case four more classes are specified with special characters $\langle b \rangle$ - the beginning of the sentence; $\langle e \rangle$ is the end of the sentence; $\langle n \rangle$ is a number, ..., $\langle u \rangle$ is an unknown word. Special characters are placed each in their own separate class and cannot be moved to other classes. It is believed that other words do not fall into these classes. The size of the class dictionary is determined by the choice of the most frequency N_v words, and the upper limit of its desired size should correspond to a limit of not more than 64 Kb. The probability of the current word belonging to the $(N-1)$ class of previous words is proposed to be estimated by the following logarithmic function

$$LL_{bi}(c)=\sum_{i=1}^{N_w}\log P(w_i \setminus C(w_{i-1}))=\sum_{j=1}^{N_c}\sum_{i=1}^{N_v}N(c_j,w_i)\cdot\log\frac{N(c_j,w_i)}{N(c_j)}. \qquad (20)$$

After some simplifications, it will be written as

$$LL_{bi}(c)=\sum_{j=1}^{N_c}\sum_{i=1}^{N_v}N(c_j,w_i)\cdot\log N(c_j,w_i)-\sum_{i=1}^{N_c}N(c_j)\cdot\log N(c_j). \qquad (21)$$

Now consider the principle of building search procedures.

V. ALGORITHM FOR INCREASING INFORMATION RELIABILITY WITH THE MECHANISM FOR CORRECTION OF DETECTED ERRORS

The general scheme for correcting distorted elements (words) of a document can be described in the form of the following steps: the first word of the concept of the document (sentence) is searched in the class of the most frequent words; if it is not there, then the search is carried out in the class (classes) of words that may be in the first place of sentences; if it is not here either, then either a refusal is issued or a search is made in the full dictionary of word forms; subsequent words are initially searched for in word classes that may follow the previous, etc.

The algorithm is tested on the basis of pseudo-grammar classes of words and frequency dictionaries of n-grams of the context of materials. For this, a rather extensive corpus of grammatically correct words of a popular literary text from the literature of 16 volumes is presented. In experimental studies, they used a corpus of approximately 0.5 million words, in which there were about 65 thousand different words.

These words with the help of a written program in the Java language are divided into 20, 50, 100, 200, 500 classes and the results are used to recognize elements of a given test text.It is determined that the frequency of the n-th word in such a list will be approximately inversely proportional to its ordinal number n, i.e. the rank of this word, i.e. the first word in use is found about two times more often than the second, and three times more often, than the third, etc. A methodology for applying the George Zipf law was developed and implemented, which made it possible to model the probability processes of the distribution of the frequency of occurrence of words in the studied context of materials.

The studies were conducted in five cases with a total volume of 404637 words. Of these, the first corpus contains 79893 words; the second body - 96157 words; the third - 97369 words; the fourth - 64841 words; fifth - 66377 words. 135 points of the word form dictionary are necessary and sufficient to take into account half of the corpus, which proves the sufficiency of the generated volume of the word form dictionary. The averaged frequency characteristics of k-grams in the tested text boxes are compared to the Zipf distribution laws with a guaranteed probability of $\Re = 0{,}95$.

Compiled 10 sets of document elements, each of which is 10^5 ten. sign. In 7 cases, confirmation of the Zipf distribution hypothesis is observed.

VI. CONCLUSION

An EDMS was developed and applied, a software package for detecting and correcting multiple information errors, which was tested in the Sphinx-4 framework.

The complex consists of an “interface” module, which is supported by MFCC, PLP and LPC and mechanisms for extracting statistical, dynamic and specific characteristics of document elements; linguist module supported by various language models, including CFGs, FSTs and n-grams; “decoder” module, supported by Viterbi or Bushderby tools and including a parallel search mechanism; module "Configurator", which serves to optimize the processing of documents.

REFERENCES

[1] Keldysh N. V. Metodicheskie osnovy avtomatizirovannogo resheniya zadach vedomst¬vennogo elektronnogo dokumentooborota. // Nauch.metod. sbornik № 56. - M., 2009. - S.110-117.
[2] Bessonov, S. V. Optimizaciya elektronnogo dokumentooborota v korporativnyh sistemah: dis. kand. ekon. nauk. M., 2000 g.,187 s.
[3]Gudov, A. M. Ob odnoj modeli optimizacii dokumentopotokov, realizuemoj pri sozdanii sis¬temy elektronnogo dokumentooborota, sbornik «Vychislitel'nye tekhnologii», 2006. - S. 53 - 65.
[4]ZHumanov I.I., Ahatov A.R. Ocenka effektivnosti programmnogo kompleksa kontrolya dostovernosti tekstovoj informacii sistem elektronnogo dokumentooborota // «Himicheskaya tekhnologiya. Kontrol' i upravlenie» - TGTU, Tashkent, 2009- № 2, s. 46-52
[5]Kemal Oflazer, Sergei Nirenburg, Marjorie McShane. Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. // To appear in Computational Linguistics, 2001.
[6]Dilek Zeynep Hakkani-Tür, Kemal Oflazer, Gökhan Tür. Statistical Morphological Disambiguation for Agglutinative Languages // Proceedings of COLING 2000, Saarbrucken, Germany, August, 2000.
[7] Baker J. K., “The Dragon system - an overview,” in IEEE Transactions on Acoustic, Speech and Signal Processing, vol. 23, no. 1, Feb. 1975, pp. 24–29.
[8] Forney G. D. “The Viterbi algorithm,” Proceedings of The IEEE, vol. 61, no. 3, pp. 268–278, 1973.
[9]Brown P., Pietra V. D., DeSouza P., Lai J. and Mercer R. Class-based n-gram models of natural language. Computational Linguistics, 1992.18:467–479.
[10]Ney H. and Kneser R. Improved clustering techniques for class-based statistical language modelling. In European Conference on Speech Communication and Technology (Eurospeech), 1993. pages 973–976, Berlin.
[11]George K. Zipf, The Psychobiology of Language, Houghton-Mifflin, New York, NY, 1935 (http://citeseer.ist.psu.edu/context/64879/0)
[12] Akhatov A.R., Jumanov I.I. Improvement of text information processing quality in documents processing systems // 2nd IEEE/IFIP International Conference In Central Asia On Internet ICI-2006, September 19-21, International Hotel Tashkent, Uzbekistan

AUTHOR’S BIOGRAPHY

Jumanov Isroil Ibragimovich

Samarkand State University,
Doctor of Technical Sciences. Professor.

Samarkand State University, 140104, Uzbekistan
s. Samarkand, University Blvd, house 15.

The author of more than 450 scientific and methodological works, including 3 monographs, 28 educational and methodological developments, 25 scientific and methodological works, 15 certificates of registration of computer programs.

Jumanov I. I graduated from Polytechnic Institute of the city of Penza in 1968.
Theme of the PhD thesis: “Methods of information control in automated control systems of mining enterprises” 1974, Institute of Cybernetics, Uzbekistan.
Theme of doctoral dissertation: “Development of the theory. methods and algorithms for monitoring information with statistical redundancy”, 1984 Institute of Cybernetics, Uzbekistan.
To date, under his scientific supervision, 17 candidates of sciences from among graduate students and applicants from SamSU have been trained in the specialty 05.13.01-System analysis, management and information processing.

Prof. I.I. Jumanov is the head of the regional scientific seminar “Information Technologies” and the organizer of a number of scientific and practical conferences of the republican and regional level devoted to the use of information and communication technologies in education. Over the past 5 years, he has prepared 3 candidates of science, supervised 10 master's theses, is the supervisor of 3 candidates for the degree of candidate of science and 2 candidates for the degree of doctor of science.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Issue 9 , September 2019

Karshiev Khusan Berkinbayevich

Samarkand State University,
Basic doctoral student

Samarkand State University, 140104, Uzbekistan
s. Samarkand, University Blvd, house 15.

Author over 20 scientific and scientific-methodical works, including 2 educational manual and 3 certificates on registration of software.

He participated in 2 international, 15 republican scientific conferences.

Karshiev Kh.B. graduated from the Applied Mathematics and Informatics Department in 2005, in 2007 received a Master of Science degree in Applied Mathematics and Information Technology.

In the field of scientific activity is engaged in research of a new scientific direction on the development of scientific and methodological foundations, methods and algorithms for increasing the reliability of information using mechanisms for extracting logical and semantic links of elements of electronic documents

He actively participates in reports at the scientific and methodological seminar. He reads lectures, conducts practical and laboratory classes in the fields of programming, computer networks, information security.

Introduces modern systems of information and pedagogical technologies in the educational process.