UDC 004.94:658.01
V. Osypenko[1], A. Shvorov[2]
[1]National university of life and environmental sciences of Ukraine,
[2]National technical university of Ukraine "KPI"

ABOUT SOME DESIGN PRINCIPLES OF INFORMATION-RETRIEVAL SYSTEM AND PROCESSING OF ELECTRONIC DOCUMENTS IN INTERNET

The paper analyzes the main trends of search the basic direction and identifying the priority of electronic documents handling in the Internet. The technical principles of construction of information retrieval system and ability to use linguistic processor and neural networks for solving problems determining the importance of consideration the input documents were grounded. The object of the study was the decision support systems designed to determine priority of review incoming documents. The proposed system can be actively used in information support blocks of inductive technology of system information-analytical research.
Keywords: electronic documents, incoming documents, linguistic processor, neural network.

Introduction

One of the complex and specific types of operator activity in a information retrieval systems (IRS) in the Internet is an activity that is related to the analysis of information flows in order to highlight and primary handling the most important the input electronic documents (IED). This raises a need to in the preceding of automatic abstracting of each IED as well as for analyzing of complex information parameters to assess their importance.

1. The analysis of recent research and purpose of papers

Our analysis reveals that one of the efficient processing ways of IED given their importance is widely used of decision support systems (DSS) to determine the sequence of processing IED [1-3]. However, currently not fully in modern literature is considered the preliminary automatic abstracting of each IEDs as well as an analysis of complex of informative parameters to assess their importance by using the linguistic processor (LP) as a component of the DSS.

The purpose of this paper is to determine the basic elements of design the search engine in the Internet and the structure of DSS in terms of determining the sequence of IED consideration in view of their importance.

2. The main material of research

In developing the IRS, which is an integrated expert system, the following main stages of its creation can be allocated: identification, conceptualization, formalization, implementation, testing, maintenance testing [3].

At the stage of identification are defined the tasks to be resolved using IRS during its operation, and turns out the goal of development as well as the necessary resources. Stage conceptualizing involves the meaningful analysis and justification of methods for solving functional tasks IRS. At the stage of formalization are determined the detection methods and interpretation of all kinds of knowledge, simulated the operation of the system, executed the assessing the degree conformity of planned and achieved goals IRS operation, assessed the adequacy of representation and manipulation of knowledge. Stage performance is characterized by obtaining of knowledge from experts.

At the stage of testing by an expert group is carried out the check of possibility of using an integrated expert system in the process of finding relevant information in the Internet. Testing continues for as long as the expert group concludes that system reached the required level of competence.

At the stage of exploitation testing is verified the suitability of IRS to solve functional problems and the degree of goal achievement. As a result of this phase you may need the substantial modification of system. The most difficult is the initial stages of IRS construction of, that requires further research.

So, on considered the stages provided justification of approaches to implementation the following basic functions of IRS: IED-search by keywords and parameters; forming the abstract of IED; determining the sequence of issuance abstracts IED to the operator depending on their importance; IRS opportunity for self-learning, as well as the convenient interaction of operator with IRS.

In order to form abstract IED there is a need for a special linguistic support ICS, based on a linguistic database (LDB), which includes various dictionaries specified format [4]. With LDB linguistic processor to solve the problem of decomposition of the text IED for these abstract components of the IED.

The sequence of IED abstracting includes the following stages:

- nomination of the previous hypotheses about the meaning of all IED;
- determining the values of obscure words (special terms);
- create a common hypothesis (about knowledge);
- clarify the meaning of terms and interpretation and interpretation of certain text fragments under the influence of the general hypothesis (from the whole to particle);
- forming semantic structure of the text IED through the installation of internal relations between key words and fragments, as well as due to the formation of abstract concepts, generalizing particular fragments of knowledge;
- adjustsing the overall hypothesis about the meaning of IED and interpretation of individual pieces of text under the influence of the general hypothesis (from particle to whole);
- adoption of basic hypotheses.

Thus, on the basis through the use of both deductive (from whole to particle) and inductive (from particle to whole) components provided the abstracting of IED.

In addition, by using LP determined the importance of IED that characterizes the content and urgency of working IED, as well as the priority of destination from which received the IED. To solve the classification problem, i.e., classifying IED, characterized by a set of indicators of how important or most important, it is proposed to use the so-called hybrid fuzzy classifier [5-7]. This classifier is a system that combines both structural and in functional terms, the principles of neural network models and fuzzy logic data processing, respectively.

The structure of the neuron consists of multipliers (synapses) an adder and nonlinear (linear) converter. Synapses provide the connection between neurons and multiple the input signal on number that characterizes the strength of connection - synapse weight. The adder unites signal fed by the synaptic connections from other neurons and external inputs. Nonlinear (linear) converter realizes the function of one argument - the output of the adder. It is called as "activation function" or "transfer function" of the neuron. Neuron generally realizes the scalar function of a vector argument.

A mathematical model of a neuron is described by the relations:

$$s = \sum_{i=1}^{n} \omega_i \cdot x_i + b, \tag{1}$$

where: ω_i , $(i = 1,...,n)$ - the weight of the synapse; b - bias value; x_i - component of the input vector (input), $(i = 1,...,n)$; s - the result of the addition; y - output of neuron; n - the number of neuron inputs; f - transformation (activation function or transfer function).

Thus, the neuron is completely described by its weights ω_i and transfer function $f(s)$. After receiving a set of numbers (vector) x_i, the neuron produces a some number y on output. That is, neural networks work is to convert the input vector *X* in the output vector *Y*, and this transformation is given by network weights.

Let us consider in more details the process of functioning of the neural network to determine the importance of the IED. The neural network generates an output signal *Y* according to the input signal *X*, realizing the some functions $Y = G(X)$. Type of function *G* determined by the values of synaptic weights and network displacements. Then, let the solution of some problem is a function $Y = F(X)$ defined by pairs of input-output data $(X^1, Y^1), (X^2, Y^2), ..., (X^N, Y^N)$ for whom $Y^k = F(X^k), k = 1, 2, ..., N$. Accordingly, the study is the synthesis function *G*, which is close to *F* in the sense of a function error *E*.

If the set of training examples are selected, i.e. pairs of $(X^k, Y^k), \ k = 1, 2, ..., N$ and method of calculating of error function *E* is determined, then the training of the neural network is transformed into a multi-dimensional optimization problem, which has a very large dimension. In this case, since the function *E* can be arbitrary form, in general – it is many-extreme nonconvex optimization problem.

Classically, neural networks are used to: the classification of images, clustering (categorization), approximation, prediction/forecasting, optimization tasks, management etc. However, in view of the ability of neural networks to adequately operate under uncertainty and incompleteness of information, it is advisable to use them to acquire new knowledge about the objects.

The method of such application of neural networks shall test to recognize of importance of IED keywords.

According to the scoring method, the input network parameters are:

- density of the keyword in the title IED (*Z1*);
- weight of keywords in IED (*Z2*);
- frequency of keywords on a web page (*Z3*);
- frequency of keywords in IED (*Z4*);
- factor taking into account the location keyword (*X1*): the title page (*title*); location of keyword in meta tags; appear on the page; replaces the text in the picture, as well as in a hyperlink.
- clearance of keyword: style; font size; font fat content (*X2*);
- site redesign (*X3*);
- correction factor (*X4*).

During the study, other factors were not used. The output of the network will generate the importance of keywords divided by the expert of importance in standard units (s.u.).

To construct a mathematical model used the linear neural network with the corresponding simple analytical expression:

$$Y = X * W + B, \tag{2}$$

where: *Y* - the network output; *W* - weight ratio; *X* - input; *B* - bias.

In demos Statistica 6.1 software environment were synthesized the linear neural network IRS (Fig. 1) with characteristics: training error (RMS of error for all outputs) - 0.103727 s.u., control error - 0.151430 s.u., test error - 0.123600 s.u.

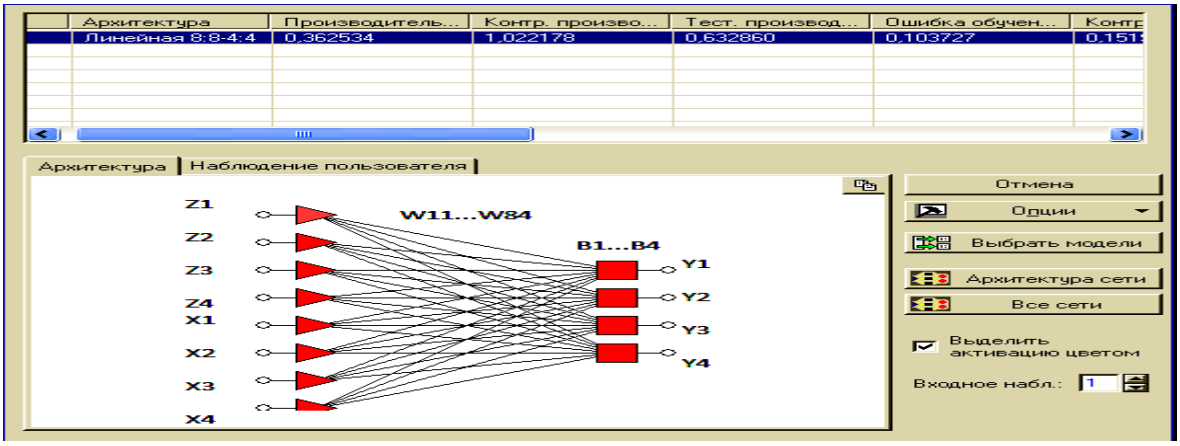

Fig. 1. Neural network architecture.

On Fig. 1: *W11 ... W84* - weights (all of them - 32, the first number – is the number of input parameters, the second - weight coefficients of output); *B1 ... B4* - biases.

Then, analytical expressions for each of the 4 outputs of the neural network are:

$$Y1 = W11*Z1 + W21*Z2 + W31*Z3 + W41*Z4 + W51*X1 + \\ + W61*X2 + W71*X3 + W81*X4 + B1; \quad (3)$$

$$Y2 = W12*Z1 + W22*Z2 + W32*Z3 + W42*Z4 + W52*X1 + \\ + W62*X2 + +W72*X3 + W82*X4 + B2; \quad (4)$$

$$Y3 = W13*Z1 + W23*Z2 + W33*Z3 + W43*Z4 + W53*X1 + \\ + W63*X2 + +W73*X3 + W83*X4 + B3; \quad (5)$$

$$Y4 = W14*Z1 + W24*Z2 + W34*Z3 + W44*Z4 + W54*X1 + \\ + W64*X2 + W74*X3 + W84*X4 + B4. \quad (6)$$

The expression of the objective function for the optimization problem in analytical form is as follows:

$$\mathrm{F} = \sqrt{(Y1-R1)^2} + \sqrt{(Y2-R2)^2} + \sqrt{(Y3-R3)^2} + \sqrt{(Y4-R4)^2} \to 0, \quad (7)$$

where: *R1 ... R4* – are the procedural values of output variables.

Given the nature of the tasks that the operator decides, it is obvious that this is the most effective organization of its activities, which provides analysis of the maximum number of the most important of IED. Efficiency of the operator of ICS will be as follows:

$$W(t_{\mathrm{a}}) = \sum_{i=1}^{n} C_i P_i(\tau_i^{\mathrm{prc}} < T^{\mathrm{avl}}), \quad (8)$$

where t_{a} - the analysis time; C_i - the importance of the information contained in IED, $\sum_{i=1}^{n} C_i = 1$; $P_i(\tau_i^{prc} < T^{avl})$ - probability of correct processing (maintenance) IED for the time that is not exceed of allowable (T^{avl}); n - number of species IED that analyzed by the operator during operation time of t_{a} [1].

The proposed performance indicator actually describes weighed probability of application maintenance IED. For a queuing system, where there is an operator position [1], the probability of maintenance is one of the most important performance indicators of IRS. It confirms correctness of the chosen performance indicator, as actually describes the weighed probability of timely analysis of all types of IED per a fixed time, in view of their importance and the average analysis time. Therefore, to take into account the importance of IED, it is necessary to find an efficient algorithm that would provide a choice from stream of VED in which there are both the greatest importance and lowest average analysis time as well as providing of the solution the operator in the form of recommendations for sequence their processing. It is possible to implement in the DSS, which automates the organization of the operator activity by means of additional information and linguistic support IRS.

The block diagram of this IPS DSS has the form shown in Fig. 2. The main elements of the system are the decision-making block (DMB) and linguistic support block (LSB). The decision-making unit evaluates the importance of IED and determined the average time of IED analysis of this type. In the operator control unit in accordance with the scheduling algorithm produced a solutions that are displayed to the operator in the form of recommendations. Block of linguistic support (LS) consists of a linguistic database that retains the details of IED dictionaries and linguistic processor that is used for forming of abstracts of IED [1].

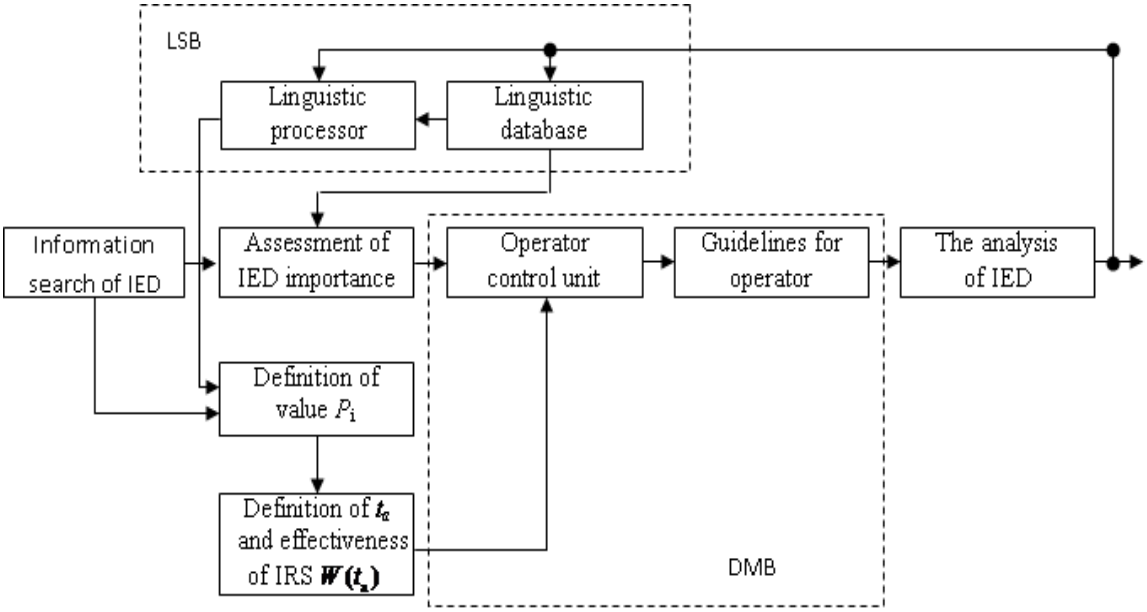

Fig. 2. Structural diagram of DSS in IRS.

How show the results of experimental studies, the application of the proposed apparatus in decision making support system enhances the efficiency of the operator to determine the sequence of processing IED in IRS 10-15%.

The described above approach can be actively used in other applications, such as inductive technologies of system information-analytical research (IT SIAR), particularly, in blocks of information-analytical support at execution of integrated innovative projects [8]. Search and processing of relevant information in such tasks could be more effective under applying of balance information bases criterion $CR_{inf}(I_b)$ [8] when using at least two analytical groups.

This criterion is responsible for the informative component in IT SIAR and is designed to select on *s*-step of inductive procedures of system research only such information I_b^{s+}, requests for which should have resemblance context of both analytical groups, closing to the current information base to optimal I_b^*. The final optimal information basis I_b^* of inductive SIAR-technology is an information basis for the creation of the final document of innovative research. Briefly remind the principle of action of balance information bases criterion $CR_{inf}(I_b)$.

At each step of the study, it compares the information requests (in practice it can be a numbered list of questions analysts in any form, questionnaire) from each groups A and B. For a single step of research:

$$CR_{inf}(I_b) = \sum_{k=1}^{K} \delta_{sk}, \quad s = 1,...,S$$
$$where: \ \delta_{sk} = \begin{cases} 0, \ if \ \ I_b^{sk(A)} = I_b^{sk(B)} \\ 1, \ if \ \ I_b^{sk(A)} \neq I_b^{sk(B)} \end{cases}, \tag{9}$$

K – is the number of allowable questions in requests on *s*-step of research and, thus, on *s*-th step the information support group offers for innovative project analytical group a some information "portion" I_b^{s+}, which minimizes the criterion (9):

$$I_b^{s+} = \arg\min_{I_b^s \in \Im_b} CR_{inf}\{I_b^{sA}, I_b^{sB}\}, \tag{10}$$

where J_b - all information requests from both groups in the *s*-th step. For the entire analytical cycle of the project:

$$CR_{inf}(I_b) = \sum_{s=1}^{S} \sum_{k=1}^{K} \delta_{sk} \to \min \ . \qquad (11)$$

Actually, selection of relevant pieces of information and could be under the responsibility of decision support systems, discussed above, that is, such system could helps to produce the relevant pieces of information in research under IT SIAR.

Conclusions

Considered the technical and methodological principles of gaining knowledge with the architecture using neural networks is advisable to apply when determining the sequence of processing documents with regard to the criteria of importance. The practical value of the proposed approach has been demonstrated. The results can be used at construction of information retrieval structures based on the calculated intelligence systems. The proposed system can be actively also used in information support blocks of inductive technology of system information-analytical research.

1. Intelligent decision support systems: theory, synthesis, performance / V.A. Tarasov, B.M. Gerasimov, I.A. Levin, V.A. Korniichuk. — K.: MAKNS, 2007. — 255 p. [In Ukraine]. 2. Gerasimov B.M. Fuzzy sets in problems of design, management and information processing // B.M. Gerasimov, G. Grabowski, N.A. Ryumshyn. — K.: Technica, 2002. — 140 p. [In Russian]. 3. Gerasimov B.M. Decision support system in automated of real-time control system / B.M. Gerasimov, V.M. Glutsky, A.A. Rabchun // Artificial Intelligence. — № 3. — 2000. — P. 39-47. [In Russian]. 4. The model of linguistic database in systems of automatic processing of natural language text data / Zamaruyeva IV, VB Tolubko etc. // Computer and math. methods in modeling. — 2013. — № 1. — P. 75-81. [In Ukraine]. 5. Shvorov A.S. Parametric method of information processing in the information-analytical systems / A.S. Shvorov // Proceedings of the Military Institute of National Taras Shevchenko University. — 2013. — Vol. 43. — P. 128-133. [In Ukraine]. 6. Rutkovska D. Neural networks, genetic algorithms and fuzzy systems / D. Rutkovska, Pilinsky M., L. Rutkowski. — M.: Hotline — Telecom, 2004. — 452 p. [In Russian]. 7. Melin P., Urias J., Solano D., Soto M., Lopez M., Castillo O., Voice Recognition with Neural Networks, Type-2 Fuzzy Logic and Genetic Algorithms. Engineering Letters, 13:2, 2006. 8. Osypenko V.V. System of criteria in inductive procedures of system-information-analytical researches / V.V. Osypenko // System technologies. — No.6 (71). — Dnipropetrovs'k. — 2011. — Pp. 106-113. [In Ukraine].