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"

ESTIMATION OF METHODS EFFICIENCY OF SEQUENCE PROCESSING DETERMINATION OF ELECTRONIC DOCUMENTS IN INTERNET

The article describes the approach to defining the importance of incoming documents and the technique of evaluating the effectiveness of application of parametric processing of electronic documents with the Internet. The assessment of the effectiveness of methods for priority processing of incoming documents in accordance with their importance.

Keywords: electronic document, priority treatment, evaluation, decision support system.

Introduction

At the time when sharply increasing the supply of incoming documents (ID) of the Internet network, and time costs required for information processing, remain unchanged, there is a need in the task priority of incoming documents based on assessment of their importance. Thus the problem of determining the importance of ID for the purpose of processing priority is given to the top spot in modern search engines.

Analysis of the set of informative parameters by which possibly determine the importance of ID [1], [2], indicates the presence of a large number of dissimilar values measured in ordinal scale, the intervals scale as well as relations and absolute scales. The vast majority of existing processing methods experimentally obtained information is not designed to account for diversity values measured parameters.

The purpose of this paper is to develop a system of evaluating the effectiveness of the method for determining the parametric processing the consideration sequence of Internet ID in view of their importance.

The presentation of research results

Typically, the processing of ID from Internet used the consistent manner view of each document without its importance. However, the storage system of full text documents require large amounts of memory and spending time on the transfer of these texts to the computer. In addition, the review process by PC-operator will last long enough. Therefore, to eliminate these deficiencies should ensure priority consideration of most important ID.

As practice shows, the importance of ID cannot be described as a simple variable flow in the separate functions or processes. This is a complex system quantity characterizing the content and working urgency of ID and recipient priority of sender.

Under the term "system" we mean the classical semantics, that is the totality (or complex) of while interacting with each other elementary structures or processes combined into one common solution of task that cannot be separately performed by any of the system components of [3, 4, 5].

Thus, the importance of ID changes over time and cannot be described by a set of fixed parameters (such as keywords, for example) for the following reasons:

- different spelling of the same words;
- among the keywords there are a lot synonyms and homonyms;
- keywords do not determine genitive-aspectual relationship between words.

The separate processing of measured parameters of ID importance and criteria of its evaluation does not always allow to give an unambiguous estimate [5, 6]. ID multilevel importance of Internet specifies the different directions for its determination.

One of these directions is to find the responses of established configurations for different behavioral situations that are the basis for selection of integrated indicators to assess the importance of ID.

In the general case under the integral method one should understand the methods based on combining of several methods for determining the certain parameters or a small homogeneous set of indicators for output on the basis of conditional ("artificial") importance estimates of ID.

The variety of incoming documents functions and a large number of factors that affect the importance of ID, is the main difficulty in solution of the task evaluating the effectiveness of different methods of determining the sequence of electronic Internet-documents processing. To solve these tasks it is possible only through the use of integrated methods for determining the importance of ID.

The problems of uncertainty and multifactorial take place both within of each component of ID and at the convolution of aggregated estimates in the integral (generalized) importance indicator of ID (K_{id}). For example, the analysis of sources [1-5] for the evaluation of ID importance you can use the following indicators (parameters):

a) the set of statistical indicators $\vec{X}^F$, including:

- the density of keywords in the title and in abstract ID that characterizes the important and essential ID (DEA), the scale of measurement – the ratio;
- the density of keywords in the text that describes the most important ID (DMD), measuring scale – ratio;
- the density of keywords in the text that describes important ID (DID), measuring scale – ratio;
- the density of keywords in the text that describes not important ID (DND), the scale of measurement – the ratio;
- other statistical indicators;

b) the set of the functional parameters $\vec{Y}^P$, including:

- based on the keyword density – the class (CID), to which belongs ID, the scale of measurement – ordinal;
- motivation (M) of priority necessity of ID processing on time, the scale of measurement – ordinal;
- other indicators;

c) the set of technical indicators $\vec{Z}^E$, including:

- the importance of the subscriber (IS), which sent the ID, the scale of measurement – ordinal;
- time of receipt of IS, scale measuring – ordinal timing and other technical indicators.

Taking into account the indicated parameters of ID, the calculation flowchart of K_{id} can be represented as shown in Fig. 1.

On Fig. 1 $\vec{\rho}_{DMD/DID}$ and $\vec{\rho}_{DND/M}$ are taken into account the expert conclusions about the existence of dependencies between individual parameters of one or more groups of indicators.

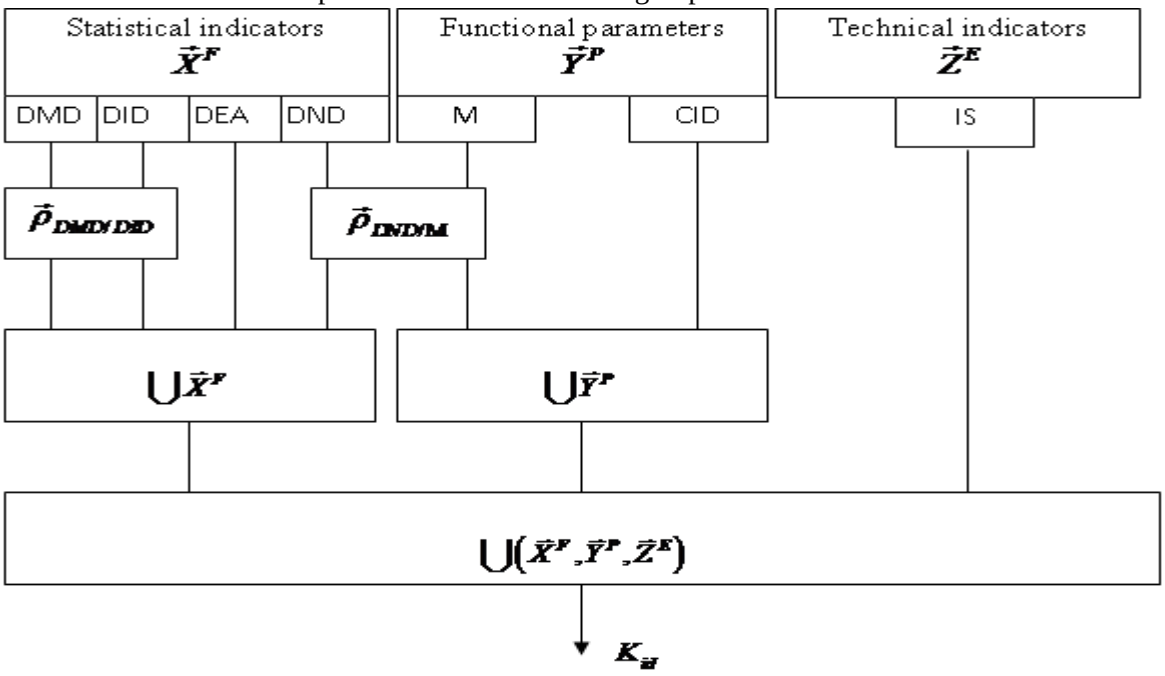

Fig. 1. Block diagram of a generalized index calculating of importance of ID

The task of this stage of technique can be solved in many ways, but in this case, stability is important final estimates experts. To such criteria may correspond estimates that are based on the median Kemeni [7].

Consider the algorithm of synthesis of such assessments, first defining a distance Kemeni interpretation, similar to [8].

In modern analytical methods for evaluation and harmonization the expert group conclusions are often used some binary relationship, including equivalence relation, the relation of similarity (tolerance) and others.

It is known that a binary relation *A* on finite set $G = \{g_1, g_2, \ldots, g_k\}$ – is a subset of $G^2 = \{(g_i, g_j), i, j = 1, 2, \ldots, k\}$. This pair is in *A* if and only if when between g_i i g_j exists such attitude.

Let $A_1, A_2, \ldots, A_s$ – random tolerances (answers), which reflect the conclusions of *s* experts. It is known, that every binary relation $A = \|a_{ij}\|$ can be represented as a matrix of 0 and 1, where $a(i, j) = 1$ when g_i and g_j are in certain relation to *A*, but also $a(i, j) = 0$ – when this relationship does not exist. Kemeni distance between binary relation $A = \|a_{ij}\|$ and $B = \|b_{ij}\|$ is called the value:

$$d(A, B) = \Sigma\left|a(i, j) - b(i, j)\right|, \quad \text{i, j} = 1,...,\text{s}. \tag{1}$$

Thus, the distance Kemeni between binary relation – is the sum of modules of differences of elements that are on the same locations and in relevant matrices or simpler – it is the number of does not coincide elements in matrices *A* and *B*. By definition [8] the Kemeni's median corresponds to the value:

$$\tilde{A}_m = \arg\min_A \sum_{k=1}^{s} d(A_k, A). \tag{2}$$

In the quoted above work indicated, and it is obviously, that the Kemeni's median is partial case of empirical average for data of non numerical nature and for it is just law of large numbers at growth of *s*:

$$\arg\min_A \sum_{k=1}^{s} d(A_k, A) \xrightarrow{s\to\infty} \arg\min_A \mathrm{M}\left[d(A_1, A)\right], \tag{3}$$

where *M* – the mathematical expectation operator.

Given this, it can be important for us to conclude that the Kemeni's median should be used as a stable benchmark assessment of the expert committee conclusions. This property the Kemeni's median we could use in the following algorithm.

Let the dependence levels between individual parameters are grouped in square symmetrical relatively main diagonal matrix, where *n* – number of indicators. These levels necessary to evaluate.

Step 1. Formation of questionnaires to the experts about the existence of relationships between individual indicators lodged above in paragraphs (a) - (c).

Step 2. Execution of expert session and receive matrices, where *s* – the number of experts.

Step 3. For element e_{ij} constructed the matrix of pairwise Kemeni' distances (1) for the set of binary relations, which includes *s* elements. This matrix has symmetrical view relative to the main zero-diagonal:

$$\Delta = \begin{pmatrix} 0 & d_{12} & \ldots & d_{1s} \\ d_{21} & 0 & \ldots & d_{2s} \\ & \ldots & & \\ d_{s1} & d_{s2} & \ldots & 0 \end{pmatrix}. \tag{4}$$

Step 4. Finding the median of Kemeni (2). For this purpose to analysis we introduce the expression:

$$\rho_j\left(A_j\right) = \sum_{i=1}^{s}\rho\left(A_i, A_j\right),\ j = 1, \ldots, s, \tag{5}$$

where A_j, $j = 1, \ldots, s$ – current matrix on which conducted the minimization and is located the meaning $\tilde{A}_m$ at which takes place (2), i.e.:

$$\rho^*\left(A^*\right) = \min_{\rho}\{\rho_j\left(A_j\right) = \sum_{i=1}^{s}\rho\left(A_i, A_j\right),\ j = 1, \ldots, s\}, \tag{6}$$

Step 5. Value of the Kemeni's median $\rho^*\left(A\right)$ assigned to element e_{ij}.

Step 6. Go to step 3.

The procedure is completed after determining of all elements of matrix $E = (e_{ij})$.

To solve the problem of classification, that is assigning ID that characterized by a set of indicators for important or for most important, it is proposed to use a special classifier [5, 6]. This classifier is a system that unites in structural and functional ways the principles of neural network models.

The task can be solved by, for example, 4-layer neural network, structural diagram of which is shown in Fig. 2.

The first layer A of the network provides at the output the degree of belonging as accordance of measured parameters ID $\{X, Y, Z\}$ to specified requirements.

The proposed version of network is designed for 3-level evaluation: "not important", "important", "the most important."

If necessary, the resolution of the classifier can be increased, which entails an increase in the number of neurons in the layer A, but it does not affect on other layers of the network and on its functioning algorithm.

The second layer B is unifying for each indicator $x_i\ (y_j\,;\, z_k)$ and is required to take into account the possibility of getting i-th signs simultaneously in two classification groups (usually with varying degrees of belonging).

The third layer C is intended to combine estimates within each group of indicators: $\vec{X}^F$, $\vec{Y}^P$ and $\vec{Z}^E$. The peculiarity connections between 3 and 4-th layers is the presence within the group ρ_{ij}^{x} and intergroup ρ_{ij}^{xy} auxiliary functional elements, reflecting the fact that the functional dependence of relevant (i, j) indicators. Introduction of auxiliary functional elements can increase the flexibility and reliability of the classifier in conditions of possible incompleteness of measurements because of time or technical problems in the normal mode.

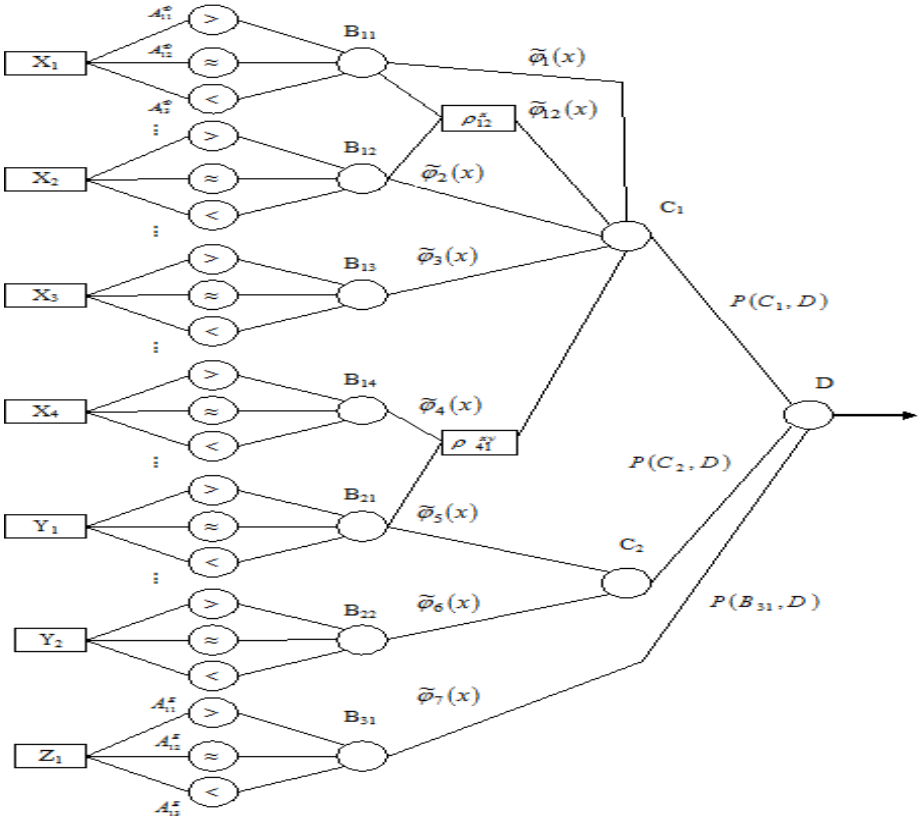

Fig. 2. Structural diagram of layer neuro-fuzzy network

Neurons of layer C is a standard neurons whose outputs are formed using sigmoid activation functions and are treated as belonging degree (measures of conformity) of statistical (functional and technical) indicators for given requirements of determine the importance of ID.

The fourth layer D is represented by a single neuron, whose inputs are weighted value of measures conformity ID for each group of indicators, and output is the measure of conformity about priority consideration of ID.

Weights of connections $P(C_i, D)$ between the third and fourth layers are determined by experts in advance depending on the specific of ID class and characterize the importance one or another group of performance.

The proposed neural network can be classified as synchronous multi-layered heterogeneous network with local connections but without feedback. It allows to remove the question on dynamic steadiness of neural network, that is an important advantage of the proposed structure.

An estimation technique of effectiveness of parametric processing methods of ID from the Internet includes the following steps:

1) determination of time cost of processing ID without priority parametric processing of ID;
2) calculation of time costs for priority processing for various methods of Internet-ID;
3) determination of most effective priority discipline of ID processing taking into account their importance.

To select the optimal organization of sequence processing of ID (on the criterion of time costs ID processing) consider a distributed computing network in terms of queuing system (QS).

The existing system of parametric processing of ID can be represented as QS with the priority queuing service discipline (FIFO).

For FIFO service discipline the average waiting time for processing materials defined as follows [9]:

$$\omega = \frac{\sum_{з=1}^{\grave{I}} \lambda_з \upsilon_з^2}{2(1-R),}, \tag{7}$$

where ω – the average waiting time of requests such i = 1, ..., M; $\lambda_з$ – intensity of i = 1,... , M; $\upsilon_з^2$ – the second initial moments of the service time of requests such i = 1, ... , M; R - total system boot $R = (\rho_1 + ... + \rho_m) < 1$.

Thus when using of service discipline with relative priority in the processing and transmission of ID, the average waiting time of requests with priorities k = 1, ... , M is defined as follows:

$$\omega_{ê} = \frac{\sum_{з=1}^{ì} \lambda_з \upsilon_з^{(2)}}{2(1-R_{k-1})(1-R_k)}, \tag{8}$$

where ω_i – the average waiting time of requests such i = 1, ... , M; k – priority of requests, k = 1, ... , M; $\lambda_з$ – intensity of i = 1, ..., M; $\upsilon_з^2$ – the second initial moments of service time requests such i = 1, ... , M; R - total system boot $R = (\rho_1 + ... + \rho_m) < 1$ [9].

The advantage of discipline with the relative priority of service is:

- time saving for ID priority compared to the FIFO;
- relatively simple implementation.

In the case where expected processing time of some ID does not satisfy the needs, compared to service discipline with relative priority, then apply the discipline of requests service with absolute priority. In this case, a priority of ID received in the system, stop processing with less priority requests. Then the average waiting time for ID with absolute priority k = 1, ... , M, provided that interrupted the less priority requests are further served from the place of interruption:

$$\omega_з = \frac{R_{k-1}\upsilon_k}{1-R_{k-1}} + \frac{\sum_{з=1}^{ì} \lambda_з \upsilon_з^{(2)}}{2(1-R_{k-1})(1-R_k)}, \tag{9}$$

ω_i – the average waiting time of requests such i = 1, ... , M; k – priority of requests, k = 1, ... , M; $\lambda_з$ – intensity of i = 1, ..., M; $\upsilon_з^2$ – the second initial moments of service time requests such i = 1, ... , M; R - total system boot $R = (\rho_1 + ... + \rho_m) < 1$; $\upsilon_з$ – mathematical expectation i = 1, ..., m [9].

The advantage of discipline absolute priority consists in reduced of service time for important ID compared with the relative priority (Fig. 3).

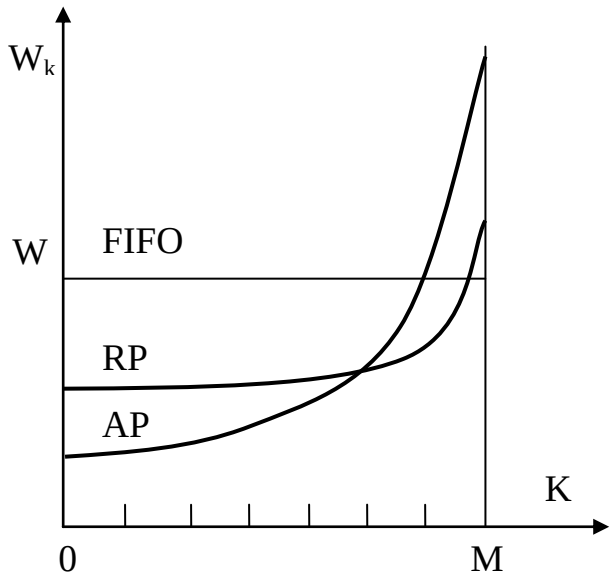

Fig. 3. The average waiting time for ID FIFO of: RP – relative- and AP - absolute priority

Thus, parametric processing of ID with absolute priority leads to a reduction of waiting time in 2-3 times for requests with high priority, but at the same time it increases the waiting time of ID of low priority. Based on this, one can say that the most effective method of parametric processing of ID with the absolute priority, that can be carried out on the basis of a special decision support system with priority consideration of Interne ID, technological principles of which have been considered in [10].

Conclusion

1. The method of parametric processing of ID to determine the sequence of their consideration of Internet taking into account their importance was described.

2. A solution for importance evaluation incoming electronic documents using a hybrid classifier, which expediently to basis for decision support systems concerning priority consideration of ID using parametric method of processing information Internet network of absolute priority have been offered.

1. Інтелектуальні системи підтримки прийняття рішень: теорія, синтез, ефективність / В.О. Тарасов, Б.М. Герасимов, І.О. Левін, В.О.Корнійчук. – К.: МАКНС, 2007. – 255 с. 2. Герасимов Б.М. Человеко-машинные системы принятия решений с элементами искусственного интеллекта / Герасимов Б.М., Тарасов В.А., Токарев И.В. – К.: Наукова думка, 1993. – 184 с. 3. Герасимов Б.М. Нечеткие множетва в задачах проектирования, управления и обработки информации / Герасимов Б.М., Грабовский Г.Г., Рюмшин Н.А. – К.: Техніка, 2002. – 140 с. 4. Герасимов Б.М. Система поддержки принятия решений в АСУ реального времени / Герасимов Б.М., Глуцкий В.М., Рабчун А.А. // Искусственный интеллект.– №3.– 2000.– С. 39-47. 5. Мельник Ю.В. Застосування взаємодіючих нейромереж в задачах визначення готовності льотних екіпажів / Ю.В. Мельник, О. Ю. Чуніхін // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – 2007.– Вип. 7. – С. 123-128. 6. Шворов А.С. Метод параметричної обробки інформації в інформаційно-аналітичних системах / А.С. Шворов // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – 2013. – № 43.– С. 128–132. 7. Кемени Дж., Снелл Дж. Кибернетическое моделирование: Некоторые приложения. - М.: Советское радио, 1972. - 192 с. 8. Орлов А.И. Нечисловая статистика. – М.: МЗ-Пресс, 2004. – 513 с. 9. Майоров С.А. Основы теории вычислительных систем. Учеб. Пособие для вузов / Майоров С.А., Новиков Г.И., Алиев Т.И. – М.: «Высш. школа», 1978. – 408 с. 10. Osypenko V. About some design principles of information-retrieval system and processing of electronic documents in Internet / Osypenko V., Shvorov A. // Вісник національного університету "Львівська політехніка". – 2014. – № 800. – С. 10–15.