UDC 81’42:519.722

USING COMPUTER TECHNOLOGIES FOR CALCULATING INFORMATION ENTROPY OF TEXTS

[1]Bikesh Ospanova, [2]Saule Kazhikenova, [1]Mukhtar Sadykov

[1]«Karaganda State Technical University», Karaganda, Kazakhstan (100027, Karaganda, Mir Boulevard, 56), e-mail: o.b.r@mail.ru

[2]«Karaganda State University n.a. E.A. Buketov», Karaganda, Kazakhstan (100028, Karaganda, Universitetskaya St., 28), e-mail: *sauleshka555@mail.ru*

In the article there are presented the results of the study the aim of which was using a measure of information amount permitting to analyze the general mechanisms of the texts information-entropic characteristics that make the basis of all spontaneously running in the surrounding world processes of information accumulation that lead to the system structure self-organization.

There is shown the result of solving a practical problem of developing the program of automated calculating the texts, whose development and realization was performed on a wide various-genre text material belonging to different styles.

Key words: algorithm, entropy, information, software, text, language, hierarchic system, self-organization, study, file.

Introductions

Informatization of the society is integrally connected with informatization of science that in turn uses information in the complex managerial, social, information-computing systems. Today computer technologies play the key role in all the spheres of social life, therefore the main type of human activity is more often methods and ways of obtaining, storing and widening knowledge and information.

Studying a language by methods of information became a prospective scientific trend from the point of view of the processes of self-organization taking place in them. Within the frames of this trend there takes place the language modeling as of a complicated, dynamic, self-organizing system from the disordered state to the ordered one.

In the article there was used the fundamental law of preserving the sum of information and entropy using Shannon’s formula. We also suggest a formula for calculating entropy and a program based on this algorithm.

Literature review

The information theory occurring is closely connected with the name of C. Shannon. The scientist introduces the concept of entropy as a measure of the knowledge uncertainty of something, and a message – as a means of the knowledge increasing. For the convenience of calculating entropy of messages transmitted by a binary code Shannon substituted the natural logarithm used by thermodynamics, $\ln$, for the logarithm with a binary base, $\log_2$:

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i , \quad (1)$$

He suggested a formula using which there can be possible to measure the information amount of the events taking place with different probability.

Thanks to this formula, scientists obtained a possibility to measure information that is contained in the code characters of quite various content. Besides, selecting logarithms as a "measure" of information, it is possible to sum up the information contained in each code character making a message and, thus, to measure the information amount contained in the whole message. Shannon's conception permitted to construct a fundamental theory that was widely recognized, practically used and continues developing intensely at present.

Information entropy is a measure of uncertainty or unpredictability of information, uncertainty of the occurring of some character of the primary alphabet. In the absence of information losses it is numerically equal to the information amount for a character of the message transmitted. [2]

Information and entropy are oppositely connected characteristics. Their mutually additional relation for any self-organizing systems is dictated by the laws of preserving the sum of information and entropy and progressive information accumulating when transiting from a lower level of organization to a higher level, up to the complete determination of the system. In this there consists the basis of comparing the texts perfection expressed by the unified information characteristic with the degree of determination of an ideal hierarchic system at each level of self-organization.

The suggested information-entropic approach to determining an objective measure of the perfection and completeness of self-organization of any texts can be considered as the development of the entropic analysis in which there is taken into account only the entropy striving to maximum. In our approach this striving is accounted jointly with the information component, besides, not in energy units but in information bits.

When characterizing the entropic-information (entropy is a measure of disorder, information is a measure of eliminating disorder) analysis of the texts, we used Shannon's statistical formula for defining the text perfection harmony:

$$H = -\sum_{i=1}^{N} p_i \log_2 p_i , \quad (1)$$

where p_i is a possibility of detecting some system unit in their set N; $\sum_{i=1}^{N} p_i = 1,\ p_i \geq 0,$ $i = 1,2,...,N$.

When determining the information amount, there is considered a language text that consists of letters, words, word combinations, sentences, etc. Each letter occurrence is described as a sequential realization of a certain system The information amount expressed by this letter is equal in its absolute value to the entropy (uncertainty) that characterized the system of possible choices and that was eliminated as a result of selecting a certain letter.

It is known that for calculating entropy it is necessary to have a complete distribution of possible combinations probabilities. That's why for calculating the entropy of this or that letter it is necessary to know each possible letter occurrence probability.

Proposed methodology

Figure 1 shows the algorithm of the proposed information entropy calculation program.

In case of combinations from two to six letters the algorithm differs in that there is made a calculation for all possible combinations.

Thus, the results of the study make suppose that any language text, form a single word to a large literary work, can be presented as a system, the elements of which are separate letters, and its parts present totalities of the same letters. So, using the synergetic theory of information, there can be carried out a structural analysis of arbitrary texts from the side of their disorder and order by the number of frequency of separate letters occurrence. We established that with transition to a higher level of organization there takes place the decreasing of the text entropy.

The studies carried out by modern science afford ground for stating that the use of a measure of information amount permit to analyze the general mechanisms of information-entropic interactions lying in the basis of all spontaneously running in the surrounding world processes of information accumulating that lead to the system structure self-organization. From this point of view it is very interesting to consider the language information interpretation. In this aspect we carried out the analysis of the existing methods of reflecting complex systems from the point of view of the entropic-information laws. The information analysis is based on the comparing of a self0organizing, abstract hierarchic system by its determinate, i.e. information component, with practical determination of various-genre texts and styles in the Russian and Kazakh languages.

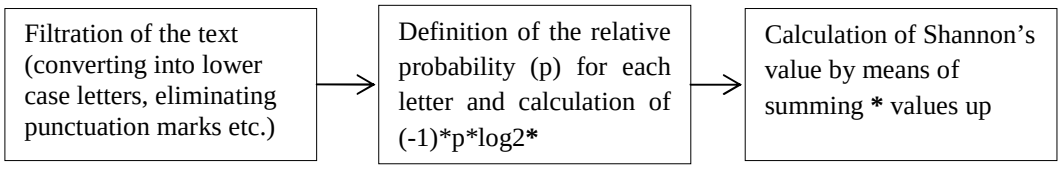

Figure 1. Raugh Flowchart for implementation of the proposed method

Software

The software interface is shown in Figure 2:

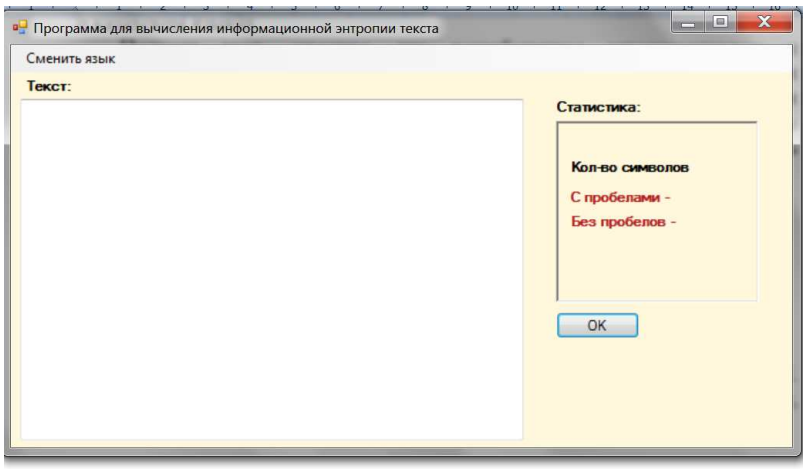

Figure 2. Software interface

After starting up there appears the window of the language selection, for which there will be carried out the information-entropic analysis. After the language selection there closes the window of the language selection and opens the window of the text information-entropic analysis. The text is introduced in the textbox field. For carrying out the analysis it is necessary to press OK.

The analysis results are recorded in the form of txt files that wiil be located in the directory of the exe file.

Conclusion

Making a conclusion, we'll note that the developed algorithmic software of the system can be used for carrying out complex studies of the texts in any language. The suggested methodology and its program realization will also permit to reduce the time consumption for calculating entropy of any text and to increase the calculation accuracy.

References

1. Shannon C. Works theory of information and cybernetics / Under ed. of R.L. Dobrushin and O.B. Lupanov. M.: Publ. of foreign literature, 1963. 827 p.

2. http://ru.wikipedia.org/wiki/Information entropy.