The method of reducing data redundancy using a randomization procedure

Ihor Lazarovych, Mykola Kuz, Mykola Kozlenko, Valerii Tkachuk, Mariia Dutchak

Vasyl Stefanyk Precarpathian National University
Ivano-Frankivsk, Ukraine

I. INTRODUCTION

Modern information systems are characterized by a large amount of operational information. In such systems, the tasks of data optimization, storage and transmission are important. Studies of typical systems for collecting information and managing of technological processes show that most of the data in such systems are weakly dynamic, and they carry redundant or unsignificant data for the user. Therefore, the task of reducing redundancy is actual.

This paper proposes a method for reducing data redundancy using a randomization procedure. The method can be effectively applied before transmitting or storing information in automation systems in various fields of communication systems and networks.

II. ANALYSIS OF RECENT RESEARCH AND PUBLICATIONS

Analysis of modern publications in the field of methods for reducing the information redundancy [1,2] allows us to conclude that their main characteristics are the compression ratio over the all range of parameter changing, as well as the complexity of the algorithm. Therefore, searching of a new methods with optimal parameters is an urgent task.

The paper [3] shows the application of the randomization procedure for noise-immune data transmission, as well for signal spectrum determination. Potential uses of randomization include reducing data redundancy.

III. PRESENTATION OF THE MATERIAL

One of the types of randomization is sorting the sequence in ascending order [3]. This is how randomization is used in the proposed redundancy reduction method.

As the practice of operating technological objects shows, the information flow of most processes is not highly dynamic, it means the value of the parameter changes slowly [4], and then remains practically unchanged for a long time (Figure 1).

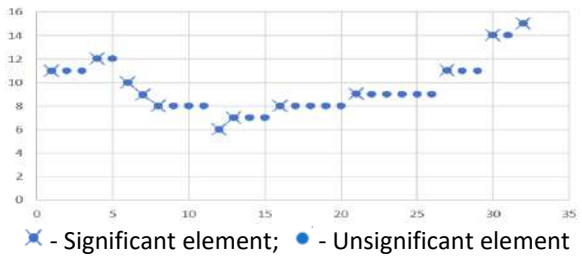

Figure 1 - Information flow of weakly dynamic process

Redundancy of information is formed due to the presence of identical values of the measured parameter.

Adaptive and non-adaptive data compression methods are widely used [5-6]. Adaptive methods are based on the analysis of the states of control objects and adaptive coding. The coding procedure is based on the identification of significant and unsignificant elements in the information sequence. When compressed, an element is considered significant if the value of the next one is not equal to the value of the current one.

It is proposed to apply a randomization procedure to arranging the data of the initial sample in ascending order, and then apply redundancy elimination by encoding only informative elements.

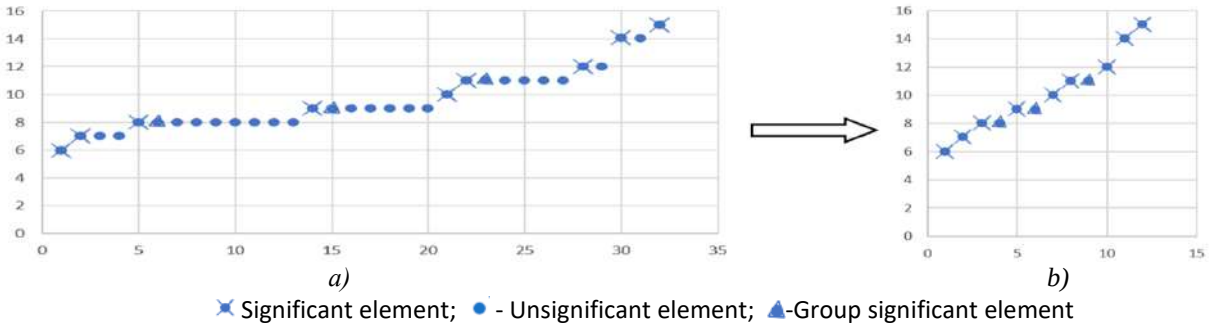

Figure 2 - Example of reducing data redundancy

Figure 2,*a* shows the dataset arranged in ascending order. The initial array contains the same informative elements. Therefore, we bring the concept of "group significant" - this is an element, the value of which is different from the following, and in addition, its value should be equal to one of the "significant" previous elements.

After arranging the sequence in ascending order, you need to encode the amplitude and sequence number of significant elements, only the sequence number of significant group elements and do not need to encode unsignificant ones. Denoting by x_a – significant, x_g – group significant, the sequence X can be written:

x_a, N_{Xa}, 0, N_{Xg}, 0, N_{Xg}, 0, N_{Xg}, ...1, x_a, N_{Xa}, 0, N_{Xg}, 0, N_{Xg}, 0, N_{Xg}, ... 1, x_a, N_{Xa}, 0, N_{Xg} ...

where N_{Xg}, N_{Xa} – accordingly sequence numbers of group significant and significant elements in the initial disordered array X*;*

*"*0*"* – means that it is followed by sequence number of the element with amplitude, the value of which follows after each new "1".

Figure 2b shows the sequence obtained after reducing the redundancy of the sequence *X* using the proposed method. In this case, the compression ratio is equal:

$$K_c = \frac{E_X \cdot n}{(n_g + n_a)(E_N + 1) + n_a E_X}, \quad (1)$$

where n_a і n_g – accordingly number of significant and group significant elements;

E_x - the number of bits representing the corresponding element *x*;

E_N - the number of bits representing the sequence number of corresponding element *x.*

The values of E_x and E_N are determined by the Hartley entropy formula [7]:

$$E_X = \hat{E}[log_2 A], \quad (2)$$

$$E_N = \hat{E}[log_2 N] \quad (3)$$

where A – quantization range, $\hat{E}$ - rounding function to a larger integer, N – the number of elements in the data frame.

On the basis of the proposed algorithm for reducing data redundancy, a software model was developed that allows evaluating the compression ratio for various types of files. A series of

experiments was carried out with 10 randomly selected files of each of listed types. The research results are shown in Table 1.

Table 1 - Results of file compression by the proposed method

File type, extension	Range of compression ratio
DOS text, *.txt	0.75-0,86
MS Word document. *.doc	1.09-1,14
application, *.exe	0.98-1.06
BMP picture, *.bmp	2.5-3.9
GIF picture, *.gif	1.25-1.57
WAV audio, *.wav	1.13-1.32
MP3 audio, *.mp3	1.04-1.29

As can be seen from Table 1, the highest value of the compression ratio is achieved for graphic files such as Windows Bitmap (* .bmp), as well as for Compu Serve GIF (* .gif) files, which already have their own compression methods. A stable compression ratio is achieved for audio uncompressed ones such as Windows Media File (* .wav) and compressed files with high-performance MP3 algorithms. The negative compression ratio for DOS text files is explained by the fact that the proposed method is effective for information in which the elements are repeated in groups. Because the text has virtually no repetition of the same characters in a row, this method is not effective.

IV. CONCLUSIONS

Thus, the proposed method for reducing data redundancy is effective and can be applied both independently and in combination with other lossless compression methods. Also, the method is especially effective for information coming from sensors and other information sources in industry and manufacturing, where processes have low dynamics. The application of the proposed method together with the well-known standard compression methods makes it possible to further reduce the amount of final information for storage or transmission.

REFERENCES

[1] A. Hanumanthaiah, A. Gopinath, C. Arun, B. Hariharan and R. Murugan, "Comparison of Lossless Data Compression Techniques in Low-Cost Low-Power (LCLP) IoT Systems," *2019 9th International Symposium on Embedded Computing and System Design (ISED)*, Kollam, India, 2019, pp. 1-5, doi: 10.1109/ISED48680.2019.9096229.

[2] Gopinath, Athira, i M. Ravisankar. "Comparison of Lossless Data Compression Techniques". 2020 International Conference on Inventive Computation Technologies (ICICT), IEEE, 2020, с. 628–33. DOI.org (Crossref), doi:10.1109/ICICT48043.2020.9112516.

[3] Lazarowych I.M. "Randomization – perspective direction of digital data processing development and constructing of special processors in computer systems". Proc. of the International Conf. TCSET‘2004., Lviv-Slavsko, Ukraine.-2004., P.403-404

[4] M. Kozlenko and M. Kuz, "Joint capturing of readouts of household power supply meters," 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET), Lviv, Ukraine, 2016, pp. 755-757, doi: 10.1109/TCSET.2016.7452172.

[5] John Edward Crosbie, Anoop Balakrishnan and others. "System and method for adaptive compression" Patent: US8949466B1, License USPTO TOS. 2015.

[6] K. Hua, H. Wang, W. Wang and S. Wu, "Adaptive Data Compression in Wireless Body Sensor Networks," 2010 13th IEEE International Conference on Computational Science and Engineering, Hong Kong, China, 2010, pp. 1-5, doi: 10.1109/CSE.2010.65

[7] S. Melnychuk, I. Lazarovych and M. Kozlenko, "Optimization of entropy estimation computing algorithm for random signals in digital communication devices," 2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), Slavske, 2018, pp. 1073-1077, doi: 10.1109/TCSET.2018.8336380