Application of information parameters for the classification of Russian-language texts

Cite as: AIP Conference Proceedings **2174**, 020123 (2019); https://doi.org/10.1063/1.5134274
Published Online: 06 December 2019

V. V. Filimonov, A. A. Zhivodyorov, Y. A. Chernykh, and L. G. Gorbich

Application of Information Parameters for the Classification of Russian-language Texts

V. V. Filimonov[1, a)], A. A. Zhivodyorov[2], Y. A. Chernykh[1, b)] and L. G. Gorbich[2]

[1]*Ural Federal University, Department of printing arts and web-design, Mira, 32, Ekaterinburg 620002 Russia*
[2]*Central Scientific Library Ural Branch of the Russian Academy of Sciences, Sofia Kovalevskaya, 22 Ekaterinburg 620137 Russia*

[a)]Corresponding author: fvv1408@list.ru
[b)]bukva32@yandex.ru

Abstract. The aim of the study is to develop a methodology for the classification of Russian-language texts. This article presents the results of assessing the applicability of information parameters for the recognition of text genres. In this work we used the results published in articles devoted to the study of the relationship of order and chaos in discrete systems. The work is performed in the department of information technologies, IRIT, UrFU

INTRODUCTION

The present work is devoted to the construction of a mathematical model of the text. At this stage, the authors are developing a technique for machine attribution of texts. The technique can be used to search for texts of the required genre and style, establish authorship and assess the usability of the text. This approach has an advantage from the point of view that allows one to analyze large amounts of information without the participation of experts and can form the basis of computer programs.

To apply the methods of machine attribution, it is necessary to use text parameters that are expressed numerically

In our previous work, we used the methods of mathematical statistics and random walk models [1]. With their help, a number of text parameters were obtained, such as: $\chi 2$ [2] statistics value, text diffusion coefficient [3], file compression percentage with WinRAR Archiver [4], frequency of each vowel in the text, etc.

The said parameters satisfy the requirements of objectivity and the possibility of numerical representation and mathematical processing of the results. These requirements were formulated in [5–9].

In [8], we obtained a compact and sufficient set of parameters, which gives the correct classification of texts with a probability of more than 80%. The set includes the value of $\chi 2$ statistics, a value similar to the diffusion coefficient in physical systems (D), the degree of compressibility of the text by the WinRAR archiver (%), and the frequencies of the letters "O" and "E" (ω_o, $\omega_э$).

In the present study, we examined the applicability of the parameters of the measure of chaos and order in discrete systems for the classification of Russian-language texts. These parameters proposed by V. B. Vyatkin in the works [10] and [11].

Classification of texts according to directions and genres is an urgent task for many areas of science, including those related to the study of stylistic features, dialects, attribution, etc.

WORK PROGRESS

The concept of order and chaos are considered and defined in the synergetic theory of information ([10, 11]). In particular, it was found that in the reflection of discrete systems through a set of its parts, the reflected information (I_0) is divided into reflected and non-reflected parts, equal, respectively, to additive negentropy (I_Σ) and reflection entropy (S). The relevant formulas are:

Physics, Technologies and Innovation (PTI-2019)
AIP Conf. Proc. 2174, 020123-1–020123-5; https://doi.org/10.1063/1.5134274
Published by AIP Publishing. 978-0-7354-1921-6/$30.00

$$I_0 = log_2 M \tag{1}$$

$$\mathrm{I}_\Sigma = \sum_{\mathrm{i=1}}^{\mathrm{N}} \frac{\mathrm{m_i}}{\mathrm{M}} \log_2 \frac{\mathrm{m_i}}{\mathrm{M}} \ , \tag{2}$$

$$S = \sum_{i=1}^{N} \frac{m_i}{M} log_2 m_i \tag{3}$$

where: M is the total number of elements in the system, N is the number of parts of the system, m_i is the number of elements in the i-th part.

Additive negentropy, eq. (2), and entropy of reflection, eq. (3), are correlated to each other in such a way that the more chaotic the structure of the system is, that is, the more parts stand out in its composition and the less these parts differ from each other in the number of elements, the more the entropy of reflection and less additive negentropy. At the same time, the greater the order in the structure of the system, that is, the fewer parts in its composition and the more they differ in the number of elements, the greater the additive negentropy and the smaller the entropy of reflection.

According to the definition, additive negentropy and entropy of reflection in the sum are equal to information (I_0), which is reflected by the system about itself.

$$\mathrm{I}_\Sigma + \mathrm{S} = \mathrm{I}_0 \ . \tag{4}$$

Thus, it can be argued that additive negentropy and entropy of reflection are informational synergetic measures of order and chaos in structure. Eq. (4) is invariant with respect to any structural transformations of the system and in the stated context is interpreted as the law of conservation of the sum of chaos and order, that is: Order + chaos = const.

In other words, whatever we do with the system without changing the total number of elements, the sum of chaos and order in the structure of the system will always remain unchanged.

Chaotic nature and orderliness in totality determine the overall structural organization of the system and, accordingly, one or another function can be used for its quantitative characterization, the arguments of which are measures of chaos and order. As such a function, V. B. Vyatkin [11] suggests using the so-called R-function (Rf), which is the ratio of additive negentropy to reflection entropy:

$$\mathrm{Rf} = \frac{\mathrm{I}_\Sigma}{\mathrm{S}} = \frac{\mathrm{order}}{\mathrm{chaos}} \ . \tag{5}$$

That is, the values of the R-function indicate what and to what extent prevails in the structure of the system: chaos or order. So, if R > 1, then order prevails in the structure of the system, otherwise, when R < 1 - chaos. With R = 1, chaos and order balance each other, and the structural organization of the system is equilibrium.

Another characteristic of the structural organization of the system is the D-function (Df), which is the result of multiplying the additive negentropy by the entropy of reflection

. $$\mathrm{Df} = \mathrm{I}_\Sigma S \ . \tag{6}$$

The D-function takes zero values in the case of absolute chaos and absolute order and expresses “the degree of development of the system”.

This article raises the question of the applicability of these parameters for the classification of Russian texts.

The aim of order to study the suitability of order and chaos parameters for solving the problem of classifying Russian-language texts, an original program "kirov" was created, allowing to calculate the values of the functions Rf and Df for texts of various genres.

Any text can be considered as a system whose elements are letters, and parts (subsystems) are sets of identical letters. Thus, in eq. (2) and eq. (3) the number of parts of the system N will be equal to the number of letters of the

alphabet, spaces and punctuation combined, and the total number of elements M is determined by the length of text (the total number of letters in the text), m_i is the number of emergence i-th letters in the text [10–12].

$$I_\Sigma \sum_{i=1}^{N} m_i = M, i = 1,2...N \ , \tag{7}$$

In accordance with the relation, eq. (4), the R-function can also be represented as follows:

$$Rf = \frac{I_0 - S}{S} = \frac{I_0}{S} - 1 \ , \tag{8}$$

As M → ∞, the entropy of reflection, eq. (3), will oscillate around a certain value $S^*, (S^* \le S^{max} = log_2 N)$ Under these same conditions, and if all the letters of the alphabet are used, the value of syntropy (2) will always monotonously increase. That is, as M → ∞, the R-function of the text becomes dependent only on its length.

Therefore, to compare texts, it is necessary to set the threshold length of the text, at which, if all letters of the alphabet are used, the R-function will be equal to 1. We will call this length the characteristic length of text, and denote it by the symbol M*

In [13], the values of Shannon's entropy (H) are calculated from the set of relative frequencies of occurrence of various letters and spaces in Russian, English, German, French and Spanish (Table 1):

TABLE 1. Shannon entropy values (H) for texts in different languages.

language	Russian	German	English	Spanish	French
H	4.35	4.10	4.03	3.98	3.96

Given that the Shannon entropy H and the reflection entropy S in this case have the same implication and are numerically equal to each other, we obtain the following values of the characteristic length of the text M^* (tabl. 2):

TABLE 2. Characteristic length for texts in various languages.

language	Russian	German	English	Spanish	French
M^*	416	294	267	249	242

Studies of the texts were conducted on material consisting of 1,162 texts of various genres. Among them are poetry, fiction, scientific, socio-political, administrative, journalistic and religious texts, memoirs. The works of foreign writers are presented in Russian translation and double authorship is indicated: the author of the original text and the author of the translation. Religious texts are translated into modern Russian.

Number of texts of each genre: Administrative – 118; Memoirs – 41; Scientific – 116; Poetry –151; Fiction –409; Publicism – 231; Religious – 96.

The D and R functions for each text were calculated as follows.

Each text was divided into fragments containing 416 characters.

For each group of 416 characters, syntropy I_Σ was calculated using the eq. (2). Then the arithmetic average of all I_Σ was calculated.

The entropy of reflection S was calculated from eq. (4)

$$S = I_0 - I_\Sigma \ , \tag{9}$$

A database was created, which includes 1162 texts with Rf and Df values calculated for them.

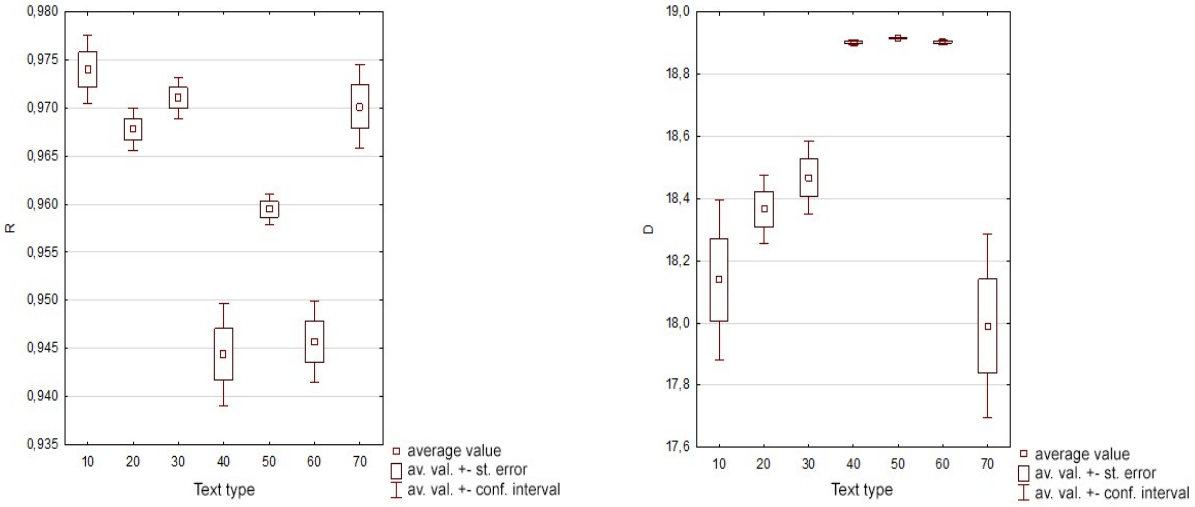
The correspondence between the genres to which the texts relate to and the mean values with confidence intervals for the R and D functions is shown in Fig. 2 and 3. The text genres are coded as follows: 10 – Administrative; 20 – Memoirs; 30 – Scientific; 40 – Poetry; 50 – Prose; 60 – Publicism, 70 – Religious.

We compared the R and D functions for texts of different genres of the array using the methods of analysis of variance (ANOVA).

CONCLUSIONS

As can be seen from the figures, the investigated texts can be statistically significantly ($p < 0.01$) divided into several groups according to the values of the R- and D-functions. At the same time one hundred percent identification of fiction (code 50) was obtained only by the value of the R-function. Thus, we can draw the following conclusions

1. According to the values of the R-function, all texts are divided into 3 groups. 1 – Administrative, Memoirs, Scientific, Religious; 2 - Artistic prose; 3 - Poetry and Journalism. The confidence intervals for all genres have a comparable size, that is, the value of the confidence interval in this case depends little on the number of texts of the corresponding genre.
2. According to the values of the D-function, two groups are distinguished. 1 - Poetry, prose, journalism; 2 - Administrative, Memoir, Scientific, Religious. Moreover, all confidence intervals in the first cluster are many times smaller than in the second. We did not find the dependence of the D-function values on the number of texts.
3. Classification of texts according to the values of R- and D-functions occurs in a significantly different way than according to the parameters we used in previous works [5–11], such as the $\chi 2$ criterion, "diffusion coefficient", etc. This means that R and D-functions carry information that affects the classification of texts, which is not available when using other parameters. It follows from this that the addition of R- and D-functions to discriminant and facto models should significantly improve the quality of recognition of a text genre.



(a) (b)

FIGURE 1 Mean values and confidence intervals R, (a), and D-functions, (b), for texts of various genres

REFERENCES

1. A. A. Kramarenko, V. V. Filimonov, A. A. Zhivodyorov and A. M. Amieva "Application of the random walk model for describing Russian-language texts" *in Information: transmission, processing, perception.* Conference proceedings (UrFU, Ekaterinburg, 2017), pp.138–164.
2. V. V. Filimonov, A. M. Amieva and A. P. Sergeev, "Clustering of Russian-language texts using χ^2 statistics" *in Information: transmission, processing, perception.* Proceedings of the International Scientific and Practical Conference (UrFU, Ekaterinburg, 2016), pp. 164–174.

3. A. M. Amieva, V. V. Filimonov and A. A. Zhivodyorov "Application of discriminant analysis to the classification of Russian-language texts". Proceedings of the 4th international conference (Editus, Ekaterinburg, 2017), pp. 65–71.
4. I. V. Selivanova, B. Ya. Ryabko and A. E. Guskov, NTI. Ser. 2. Information processes and systems **6,** 8–15 (2017).
5. A. M. Amieva, A. A. Kramarenko, V. V. Filimonov and A. A. Zhivodyorov, "Machine attribution of Russian-language texts: a review of methods" *in New information technologies in education and science.* Proccedings of the X International Scientific and Practical Conference (RGPPU, Ekaterinburg, 2017), pp. 371–375.
6. A. A. Kramarenko, V. V. Filimonov, A. A. Zhivodyorov and A. M. Amieva, "Application of the random walk model for describing Russian-language texts" *in Information: transmission, processing, perception.* Proceedings of the International Scientific and Practical Conference (UrFU, Ekaterinburg, 2017), pp. 138–164.
7. A. M. Amieva, V. V. Filimonov and A. A. Zhivodyorov, "Systematic differences statistical characteristics of texts of different genres" *in Information: transmission, processing, perception.* Proceedings of the International Scientific and Practical Conference (UrFU, Ekaterinburg, 2018), pp.140–161.
8. A. M. Amieva, V. V. Filimonov, A. A. Zhivodyorov and A. A. Kramarenko, "Statistical description of Russian texts: parameters and factors" *in Analysis of Images, Social Networks, and Texts.* Proceedings of the international scientific-practical conference (CEUR-WS, Moscow, 2017), pp. 1–8.
9. V. V. Filimonov, A. A. Zhivodyorov, A. M. Amieva and E. D. Pykhova, AIP Conf. Proc. **2015** 020022 (2018).
10. V. B. Vyatkin**,** Scientific journal of KubSAU **47**, 96–129 (2009).
11. V. B. Vyatkin, Information **10**, Iss. 4, 142 (2019).
12. V. B. Vyatkin,"Characteristic text length" *in Informatics: problems, methodology, technology*. Proceedings of the XIV International Scientific and Methodological Conference (VSU, Voronezh, 2014) **5**, pp. 263-266.
13. A. M. Yaglom and I. M. Yaglom, *Probability and Information* (Science, Moscow, 1973), pp.1-512.