Scientific Visualization, 2020, volume 12, number 4, pages 9 - 22, DOI: 10.26583/sv.12.4.02

Visual Analysis of Text Data Volume by Frequencies of Joint Use of Nouns and Adjectives

A.E. Bondarev[1,A], A.V. Bondarenko[2,B], V.A. Galaktionov[3,A]

[A] Keldysh Institute of Applied Mathematics RAS
[B] State Res. Institute of Aviation Systems (GosNIIAS)

[1] ORCID: 0000–0003–3681–5212, bond@keldysh.ru
[2] ORCID: 0000-0003-4765-6034, cod@fgosniias.ru
[3] ORCID: 0000-0001-6460-7539, vlgal@gin.keldysh.ru

Abstract

The presented research is devoted to the problems of studying the cluster structure of multidimensional data volumes. This paper presents the results of numerical experiments on the study of data volumes consisting of frequencies of joint use of adjectives and nouns. The volumes of data were obtained from samples from text collections in Russian. The aim of the research is to analyze the cluster structure of the studied volume and semantic proximity of words in clusters and subclusters. The hypothesis was used that words with similar meaning should occur in approximately the same context. In this regard, in the space of features, they will be at a relatively close distance from each other, while differing words will be at a more distant distance from each other. Research is carried out using elastic maps, which are effective tools for visual analysis of multidimensional data. The construction of elastic maps and their extensions in the space of the first three principal components makes it possible to determine the cluster structure of the studied multidimensional data volumes. The analysis of the cluster structure for the considered volume of multidimensional data is carried out. The influence of transposition of the initial data array is considered. Such analysis can be useful in the tasks of confronting negative verbal influences such as fake news, hidden propaganda, involvement in sects, verbal manipulation, etc.

Keywords: Multidimensional Data, Visual Analysis, Elastic Maps, Frequencies of Joint Use, Cluster Structures.

1. Introduction

The rapid development of the universal transition to digital technologies in the modern world has made the task of processing, visualization and analysis of multidimensional data extremely urgent. According to modern classifications, multidimensional data can be considered as Big Data. The need for processing, visualization and analysis of multidimensional data entailed the intensive development of tools for visual analytics (Visual Analytics) [1-8].

The approaches and methods of visual analytics are constantly evolving and pro-vide users with sufficiently reliable tools for solving many practical problems of re-searching multidimensional data. Such tasks include the tasks of data classification, cluster detection, identification of key determining parameters, establishing relation-ships between key parameters, etc.

In fact, the approaches of visual analytics are a synthesis of several algorithms for reducing the dimension and visual presentation of multidimensional data in manifolds of lower dimension embedded in the original volume.

Such algorithms include mapping the initial multidimensional volume in elastic maps [5–8] with different elasticity properties. These methods allow one way or an-other to separate the

cluster structure from the initial multidimensional data volume. Elastic maps turned out to be a useful and fairly universal tool, which allowed them to be applied to multidimensional data volumes of various types and different nature of origin.
This work is a continuation of research on the development of visual analytics tools for the analysis of multidimensional volumes of numerical and textual infor-mation. Studies on this topic are presented in [10-14]. In the process of research, the construction of elastic maps was tested on a large amount of data of various origins.
This work is devoted, first of all, to experiments with a multidimensional data vol-ume, which is the frequency of joint use of adjectives and nouns. With the help of certain procedures, text corpuses and arrays of frequencies of joint use are built. Earlier, in previous works, studies of a similar nature were carried out for arrays of the "verb + noun" type [10].

2. Elastic maps constructing

In this section, we give a brief description of the elastic map construction technology as a means of visualizing arbitrary multidimensional data. The ideology and implementation algorithms for building elastic maps are presented in detail in [5–8]. A description of the construction of elastic maps follows [7]. Such a map is a system of elastic springs embedded in a multidimensional data space. This approach is based on an analogy with the problems of mechanics: the principal manifold passing through the "middle" of the data can be represented as an elastic membrane or plate. The elastic map method is formulated as an optimization problem involving optimization of a given functional from the relative position of the map and data.
According to [5–8], the basis for constructing an elastic map is a two-dimensional rectangular grid G embedded in a multidimensional space that approximates the data and has adjustable elasticity properties with respect to tension and bending. The location of the grid nodes is sought as a result of solving the optimization problem of finding the minimum functional:

$$D = \frac{D_1}{|X|} + \lambda \frac{D_2}{m} + \mu \frac{D_3}{m} \rightarrow min \qquad (1)$$

where $|X|$ is the number of points in the multidimensional data volume X; m is the number of grid nodes, λ, μ are the elastic coefficients responsible for the tension and curvature of the grid, respectively; D1, D2, D3 - terms responsible for the properties of the grid.
Here D1 is a measure of the proximity of the grid nodes to the data, D2 is a measure of the extent of the grid, D3 is a measure of the curvature of the grid.
The variation of the elasticity parameters consists in constructing elastic maps with a sequential decrease in the elastic coefficients, as a result of which the map becomes softer and more flexible, adapting to the points of the initial multidimensional data volume in the most optimal way. After construction, the elastic map can be turned into a plane to observe the cluster structure in the studied data volume. On the expanded plane, you can colorize the distribution of data density on the elastic map. In some cases, such a coloring can be very useful. Elastic cards are especially effective when used in conjunction with the principal component analysis (PCA). The display of the elastic map and its sweep in the space formed by the first three principal components can dramatically improve the results, especially in clustering and classification problems. The construction of elastic maps and their scanning in the space of the first three principal components allows us to determine the cluster structure of the studied multidimensional data volumes. The author of [7] built the ViDaExpert software package [9], which allows the processing of multidimensional data volumes, the construction of elastic maps, and their effective visualization. Elastic mapping and visualization of the results in this study were performed using this software tool.
Based on the construction of elastic maps, a number of studies of various volumes of multidimensional data were carried out and a number of procedures for processing multidimensional data were developed, which significantly improved the cluster picture of the studied data volume [10-14].

3. Constructing of elastic maps for multidimensional data such as “adjective + noun”

This section presents the results of studies on the construction of elastic maps for a multidimensional data array, which are the frequencies of joint use of adjectives and nouns. This work is a continuation of [10-14], where similar studies were performed for multidimensional data volumes constructed on the basis of the “verb + noun” principle. To construct the data volume, procedures similar to those described in [10] were used. The same basic hypothesis was used that words that are close in meaning should occur in approximately the same context. In this regard, in the space of features such words will be at a relatively close distance from each other, while different words will be at a distance more distant from each other. The models “adjective + noun” were investigated. The number of adjectives was considered as the number of dimensions. The number of nouns was considered as the number of points in multidimensional space. The coordinates of these points in the space thus formed were the frequencies of joint use. In the studies considered below, the basis of the array was a sample of such frequencies for 300 adjectives and 300 nouns. That is, in this case we are considering 300 points, each of which lies in a 300-dimensional space.

The filtering procedure was carried out at the data preparation stage. Similarly to [10], to cut off the noise, all combinations with a frequency of occurrence below a predetermined frequency were discarded. In addition, only those main words (and their corresponding combinations) were selected for which the power of the set of dependent words exceeds a certain threshold value. This is necessary to filter out the noise in the combinations extracted from the collection. The threshold value of the frequency of occurrence allows us to get rid of combinations that accidentally fell into the database; the number of different combinations guarantees us sufficient statistics for comparisons.

According to the data obtained, elastic maps were constructed with a variation of the bending and tensile coefficients towards maximum “softness”. Let's consider some results.

Fig. 1 shows a fragment of the constructed elastic map for a multidimensional data array representing the frequencies of joint use of 300 nouns and 300 adjectives. A fragment of the map is presented in annotated form, showing nouns corresponding to each point.

The following figure (Fig.2) shows an extension of an elastic map in the space of the first two principal components with a coloring according to the data density. The density range is divided into five equal parts, which correspond to the colors in as-cending order from blue to red. A similar coloring is used in Fig. 2 - 8.

The presented visual image of a multidimensional array consisting of joint use fre-quencies for 300 adjectives and 300 nouns allows one to see 5 areas of condensation. Three areas are located on the left edge of the extended map, one is located in the upper right corner and another weakly expressed area of condensation is located in the lower left corner of the constructed image.

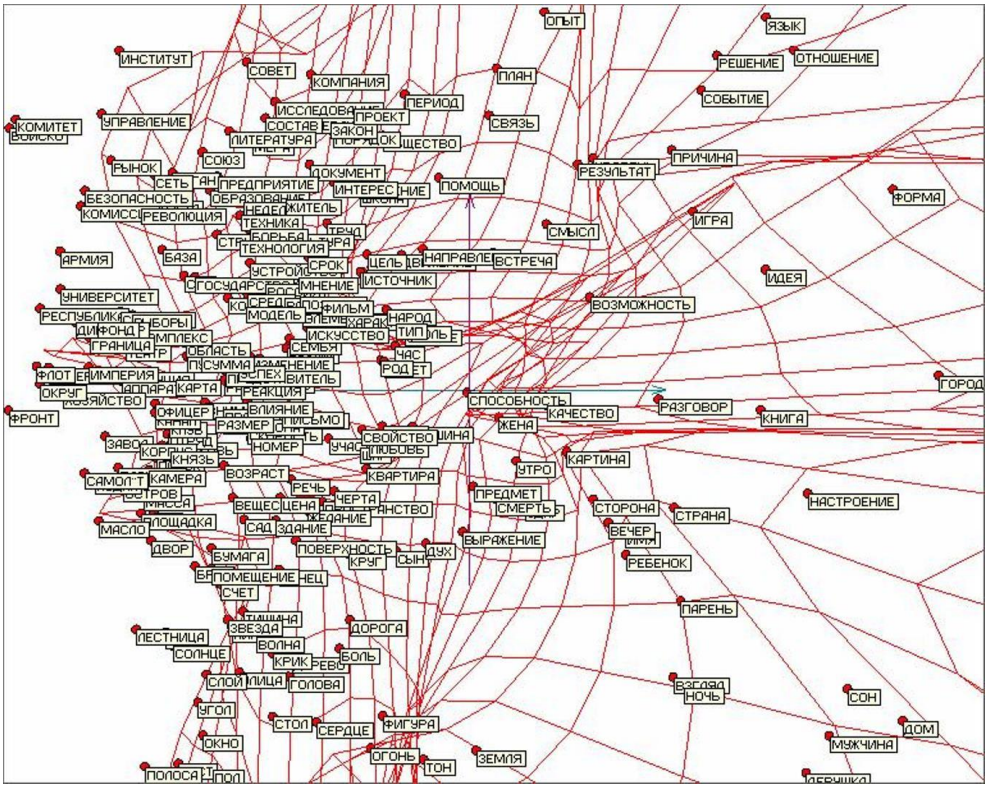

Fig. 1. A fragment of the constructed elastic map for the considered multidimensional data array with annotations

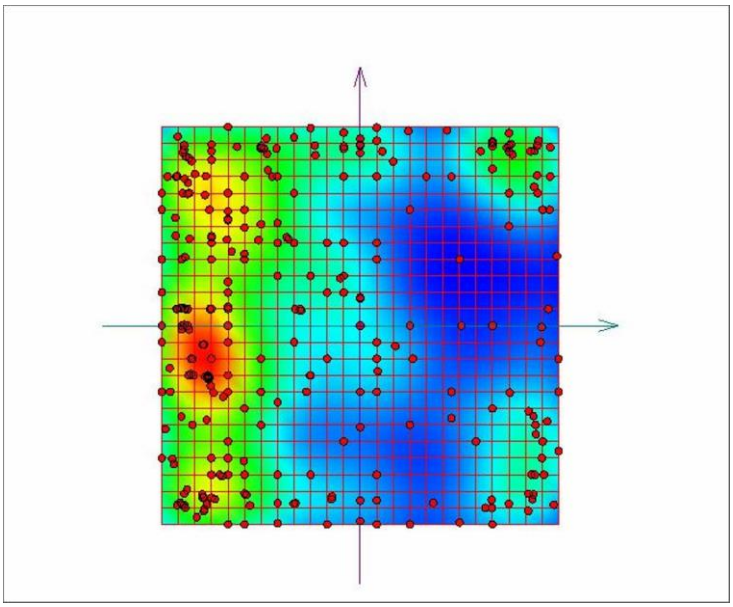

Fig. 2. Extension of the constructed elastic map for the considered multidimensional data array with coloring by data density.

Let's take a closer look at them separately.

Figure 3 shows a close-up of map fragment corresponding to the upper right cor-ner.

Here we can see a number of subclusters containing nouns that are similar in meaning. So, for example, in the left part of Fig. 3, one can trace closely related nouns ВОПРОС (QUES-TION), ПРОБЛЕМА (PROBLEM), ЗАДАЧА (TASK). Another group, located in the middle of the picture, contains semantically close nouns: МОМЕНТ (MOMENT), ПОЛОЖЕНИЕ (STATE), СИТУАЦИЯ (SITUATION), ДЕЛО (CASE), УСЛОВИЕ (CONDITION). The left-most subcluster in Figure 3 contains the words ВРЕМЯ (TIME), ГОД (YEAR), ЖИЗНЬ (LIFE), ДЕНЬ (DAY), МИР (PEACE).

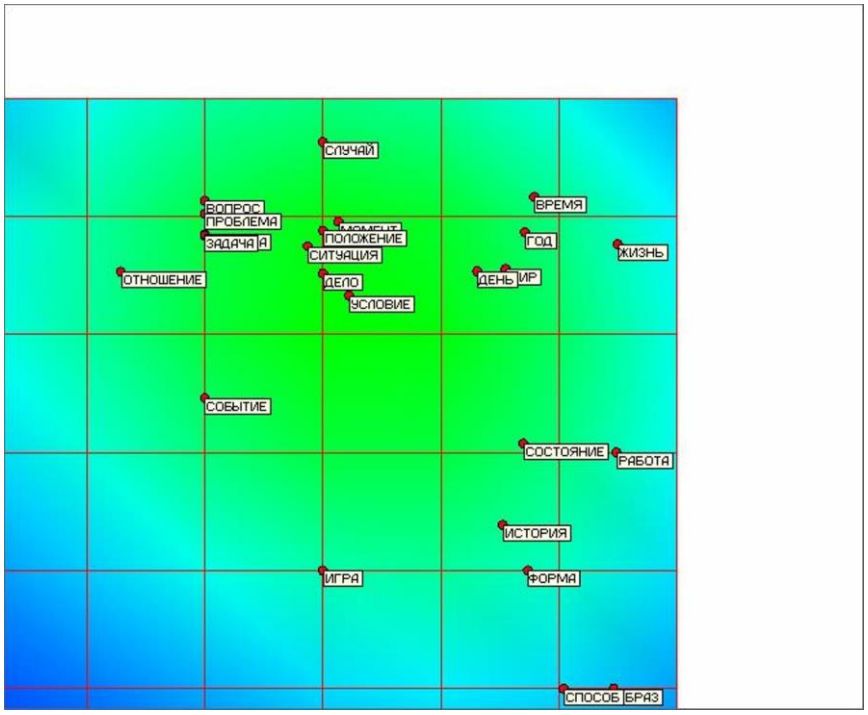

Fig. 3. Extension of the constructed elastic map - large plan - upper right corner.

Fig. 4 shows a similarly close-up map fragment corresponding to the lower upper corner. Here, judging by the density coloring in Fig. 2, there is also a weakly expressed cluster.
Here we can also see a number of subclusters containing nouns that are similar in meaning. So, for example, in the upper part of Fig. 4, we can trace the nouns МЫСЛЬ (THINK), ИДЕЯ (IDEA), close in meaning and close on the elastic map. The other group contains similar semantic nouns: ПАРЕНЬ (GUY), МУЖЧИНА (MAN), ДЕВУШКА (GIRL), ЖЕНЩИНА (WOMAN). A little lower in Fig. 4 one can see a subcluster consisting of the words ЛИЦО (FACE), РУКА (HAND), ГЛАЗ (EYE).

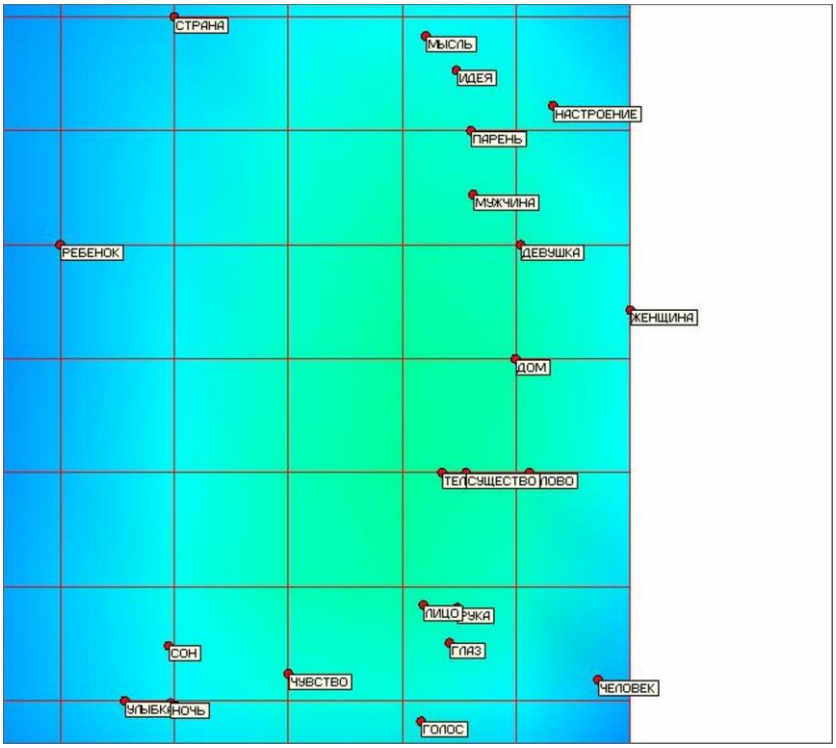

Fig. 4. Extension of the constructed elastic map - large plan - lower right corner.

A complete presentation of all the resulting clusters and subclusters in the resulting picture for nouns takes up too much space, so in the following presentation we will restrict ourselves to the most characteristic places. So, Fig. 5 shows the top-most area of condensation on the left edge in close-up. Here one can trace the following groups of concepts closely located on the fragment of the sweep. In the upper left corner there is a group of words – АРМИЯ (ARMY), ВОЙСКО (VOYSKO), ПРАВИТЕЛЬСТВО (GOVERNMENT), КОМПАНИЯ (COMPANY), ВЛАСТЬ (AUTHORITY), and НАРОД (PEOPLE). In the upper right part of Fig. 5, we can see the group ПРОГРАММА (PROGRAM), ТЕХНИКА (TECHNICS), ИССЛЕДОВАНИЕ (RESEARCH), ПРОЕКТ (PROJECT), ЗАДАЧА (TASK). In the middle on the left side of the figure is the КОМИТЕТ (COMMITTEE), РЫНОК (MARKET), РЕГИОН (REGION), УПРАВЛЕНИЕ (MANAGEMENT), ПОЛИТИКА (POLITICS) group.

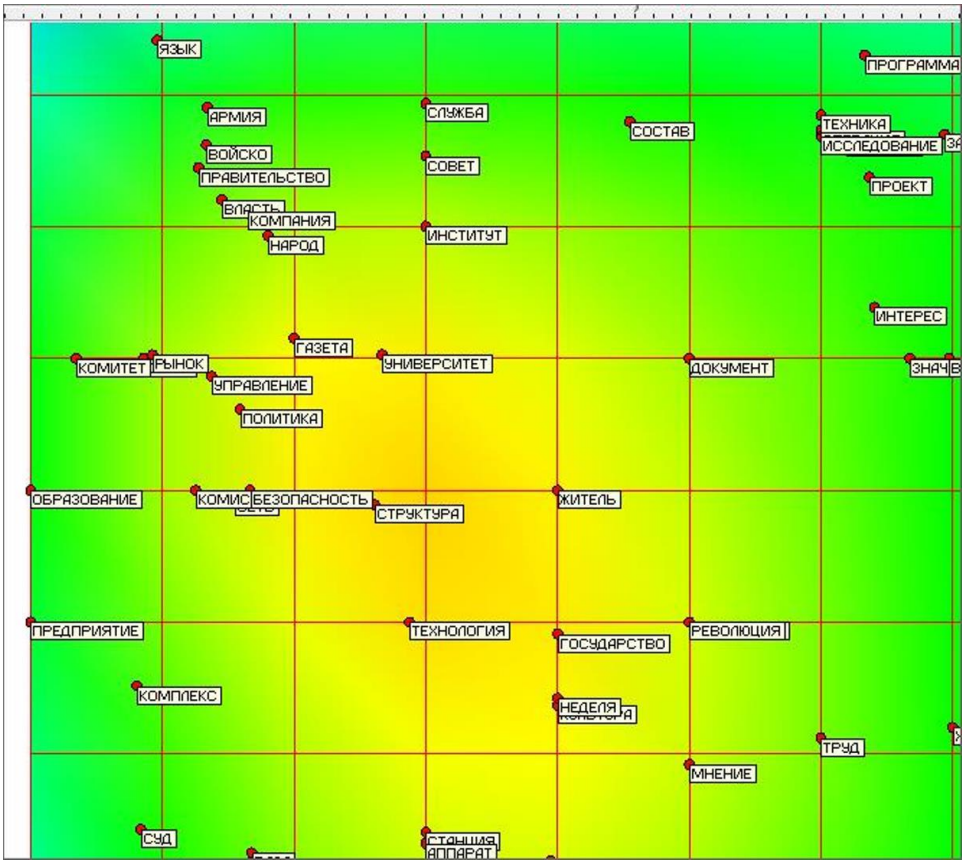

Fig. 5. Extension of the constructed elastic map - large plan - upper left corner.

Figure 6 shows the middle of the left edge of the extension. A number of distinct word groups can also be seen here. At the top of Figure 6, you can see the group - ГРАНИЦА (BORDER), НАПРАВЛЕНИЕ (DIRECTION), ОКРУГ (DISTRICT). There is a group nearby - ХОЗЯЙСТВО (FARM), РАЙОН (REGION). These two groups in terms of conceptual characteristics and location on the elastic map unfolded can be combined into a common subcluster. In the middle of the picture there is a group - ЗОНА (ZONE), УЧАСТОК (AREA). In the lower part, a group can be distinguished - ПЛОЩАДЬ (SQUARE), КНЯЗЬ (PRINCE), ОТРЯД (ORDER).

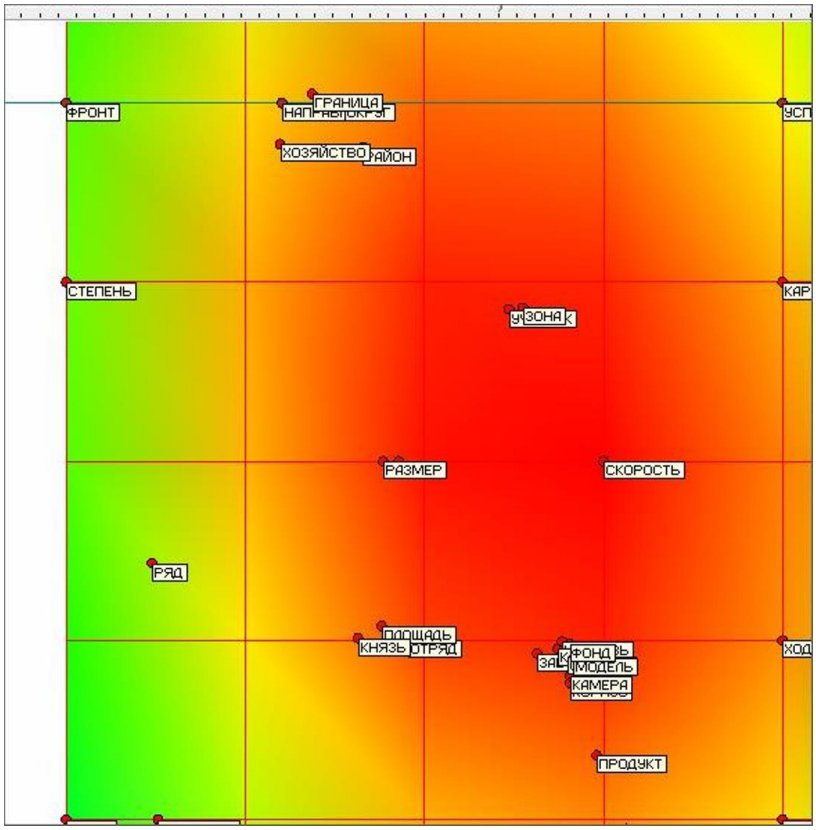

Fig. 6. Extension of the constructed elastic map - close-up image - the middle of the left edge.

The lower part of the left side of the elastic map extension is shown in Figure 7. Here the semantic groups can also be traced. In the upper left side of figure 7 - БРАТ (BROTHER), СЫН (SON). In the middle of the figure one can see the group ЛЕСТНИЦА (LADDER), УЛИЦА (STREET), ОКНО (WINDOW), КОРИДОР (CORRIDOR), ДВЕРЬ (DOOR), СТЕНА (WALL).

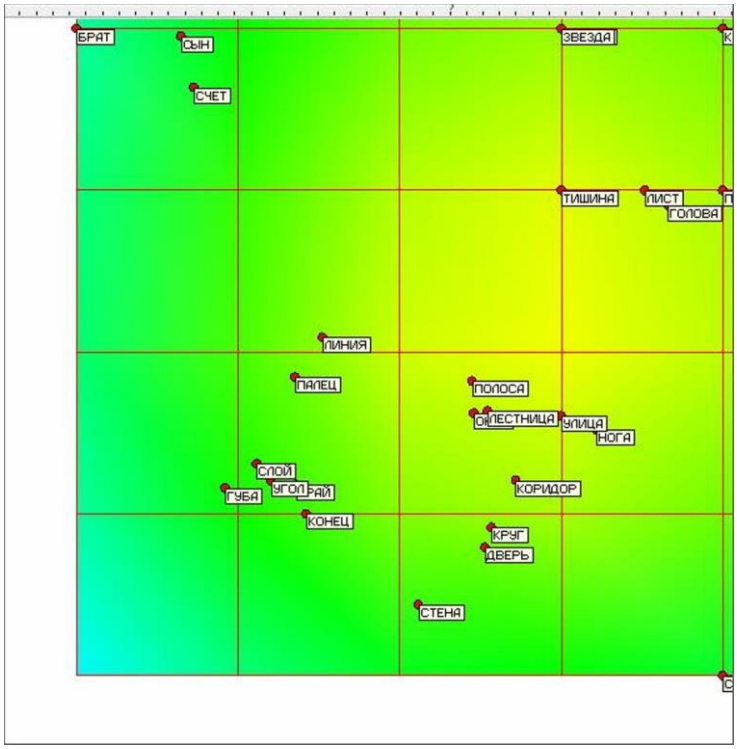

Fig. 7. Extension of the constructed elastic map - close-up image - the lower part of the left edge.

Thus, the constructed elastic map makes it possible to single out a number of subclusters and groups uniting words that are semantically related. This opens up a number of possibilities, including searching for words by related words from such group.
The considered data array was transposed similarly to [10]. We studied the transposed data array, where nouns played the role of measurements, and adjectives were considered as points in a multidimensional data array. The role of numerical characteristics is also played by the frequency of joint use of adjectives and nouns.
An extension of the constructed elastic map for colorized data density is shown in Fig. 8.
Here the picture is very similar to that shown in Fig. 2, with the difference that the weakly expressed region of condensation in the lower right corner practically disappears. The presented visual image allows one to see four areas of thickening. Three areas of thickening are located on the left edge of the map, one is located in the upper right corner.

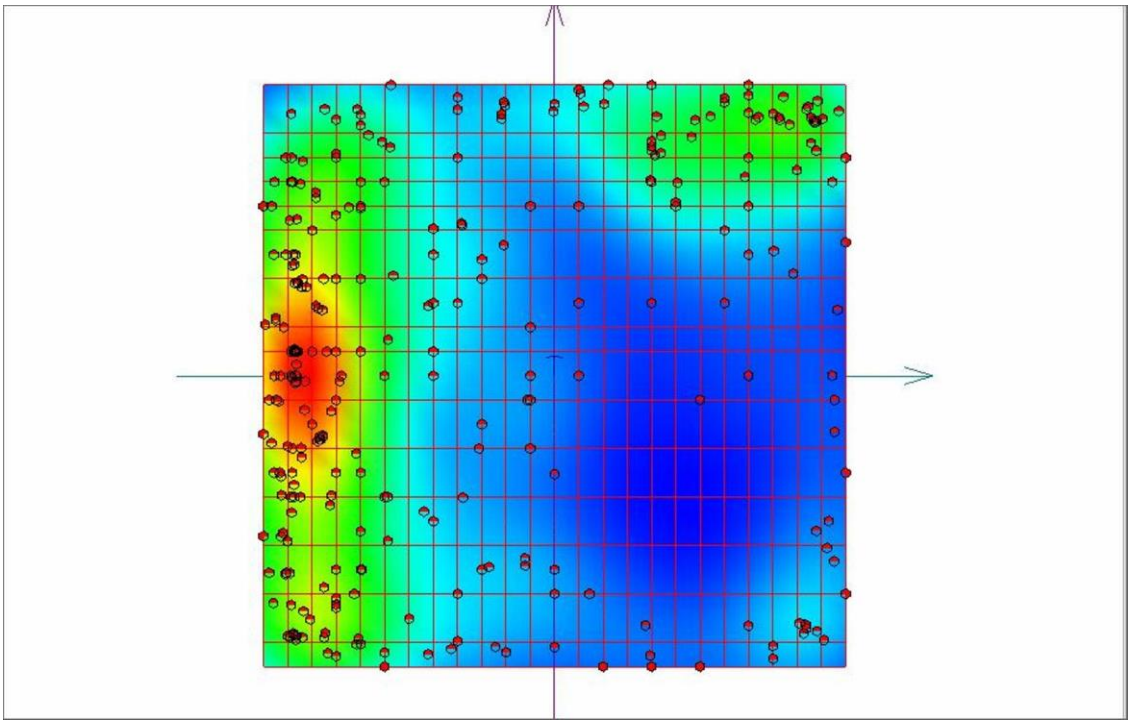

Fig. 8. Extension of the constructed elastic map for a transposed data array with coloring by data density.

As in the previous case, we consider some areas of condensation.
Fig. 9 shows a close-up of the thickening region in relation to the data density in the upper right corner of the sweep of the elastic map for the transposed data array. Here traced groups of adjectives that are similar in characteristics. For example, in the upper right corner – ГОСУДАРСТВЕННЫЙ (STATE), НАЦИОНАЛЬНЫЙ (NATIONAL), ПОЛИТИЧЕСКИЙ (POLITICAL), МЕЖДУНАРОДНЫЙ (INTERNATIONAL), ОБЩЕСТВЕННЫЙ (PUBLIC). Nearby is a group with na-tional-geographical characteristics – РУССКИЙ (RUSSIAN), ЕДИНЫЙ (UNIFIED), ЕВРОПЕЙСКИЙ (EUROPEAN), АМЕРИКАНСКИЙ (AMERICAN), ИНОСТРАННЫЙ (FOREIGN), НЕМЕЦКИЙ (GERMAN), ФРАНЦУЗСКИЙ (FRENCH), ИТАЛЬЯНСКИЙ (ITALIAN), ГЕРМАНСКИЙ (GERMAN).

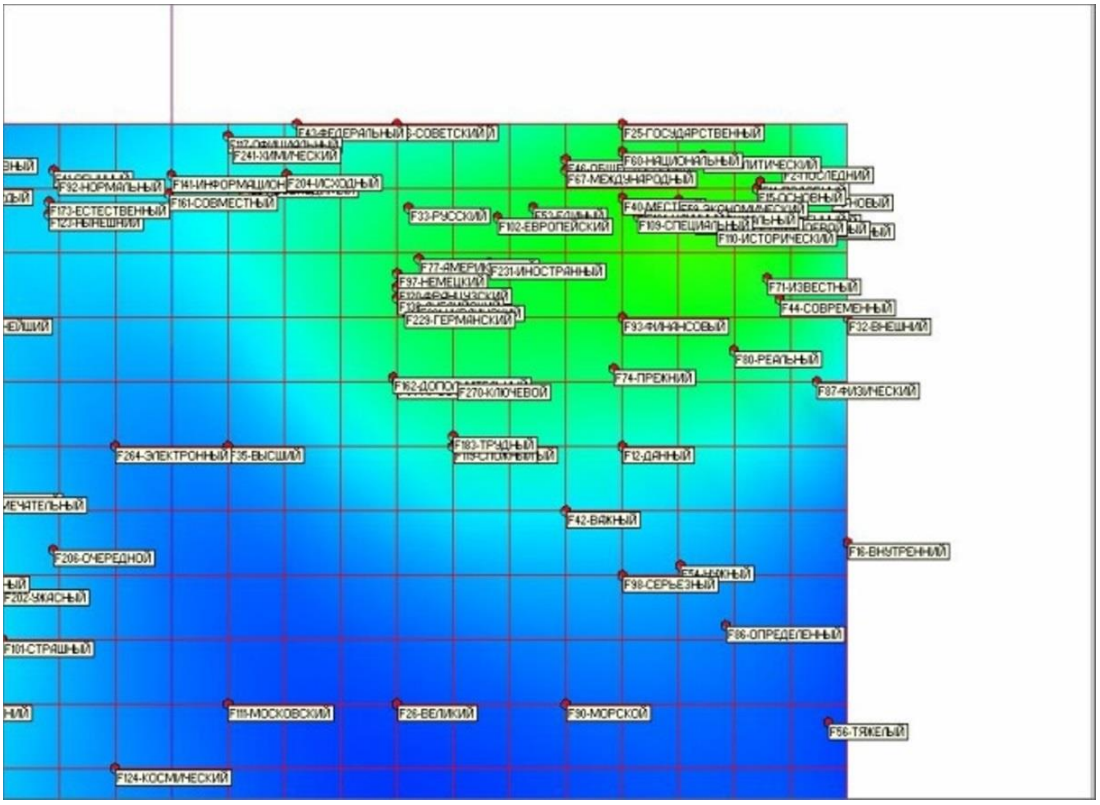

Fig. 9. Extension of the constructed elastic map - large plan - upper right corner.

We also give an example of a group of words located in the lower right corner of the constructed extension of an elastic map for a transposed array. This fragment is shown in Fig. 10. At the bottom of the figure, one can distinguish a group of adjec-tives with size characteristics – ОГРОМНЫЙ (HUGE), БОЛЬШОЙ (BIG), МАЛЕНЬКИЙ (SMALL), КРУПНЫЙ (LARGE), НЕБОЛЬШОЙ (LITTLE), МЕНЬШИЙ (LESS), БОЛЬШИЙ (LARGE), ДЛИННЫЙ (LONG), УЗКИЙ (NARROW), ШИРОКИЙ (WIDE).

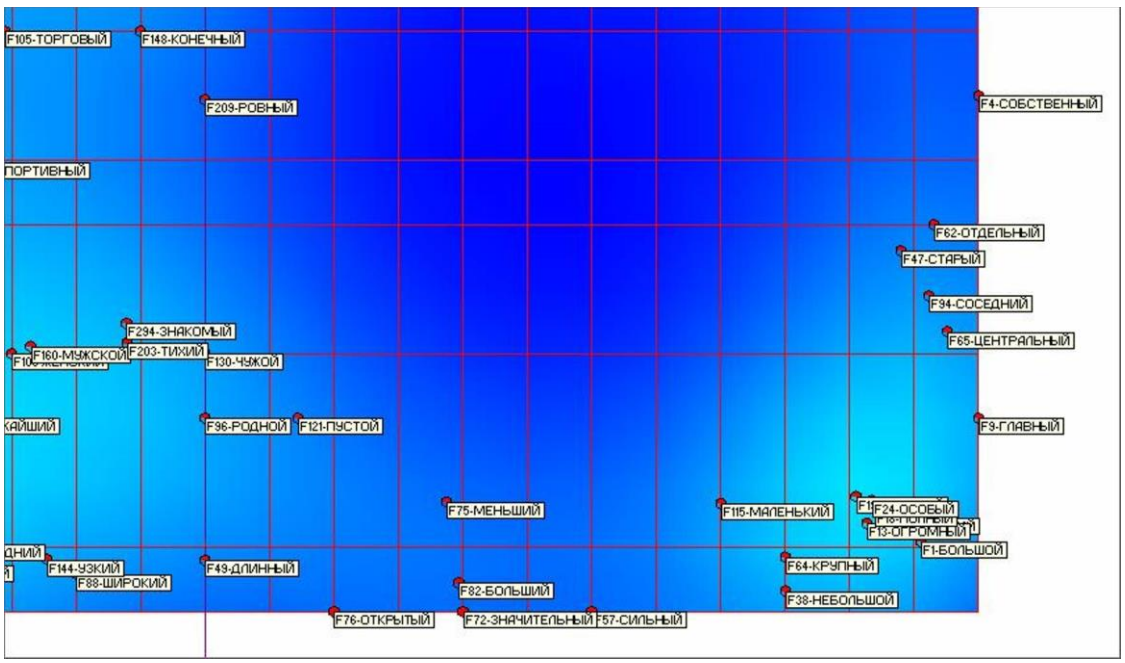

Fig. 10. Extension of the constructed elastic map - close-up - lower right corner.

Let's move on to the left side of the extension, where three areas of data condensation are located, determined by density coloring. Let's consider them one by one.

Figure 11 is a partial flat pattern showing the top left corner. A group of adjectives can be traced here - ПРОШЛЫЙ (PAST), СЛЕДУЮЩИЙ (NEXT), ПРЕДЫДУЩИЙ (PREVIOUS), АНАЛОГИЧНЫЙ (SIMILAR), СООТВЕСТВУЮЩИЙ (APPROPRIATE), ВОЗМОЖНЫЙ (POSSIBLE), ПОЛОЖИТЕЛЬНЫЙ (POSITIVE), ПРАВИЛЬНЫЙ (RIGHT), СЕКРЕТНЫЙ (SECRET). Below - СЛЕДСТВЕННЫЙ (INVESTIGATIVE), СУДЕБНЫЙ (JUDICIAL). Even lower - ПРАВООХРАНИТЕЛЬНЫЙ (LAW ENFORCEMENT), УГОЛОВНЫЙ (CRIMINAL). A group is located nearby - РЕЗКИЙ (SHARP), ГРОМКИЙ (LOUD).

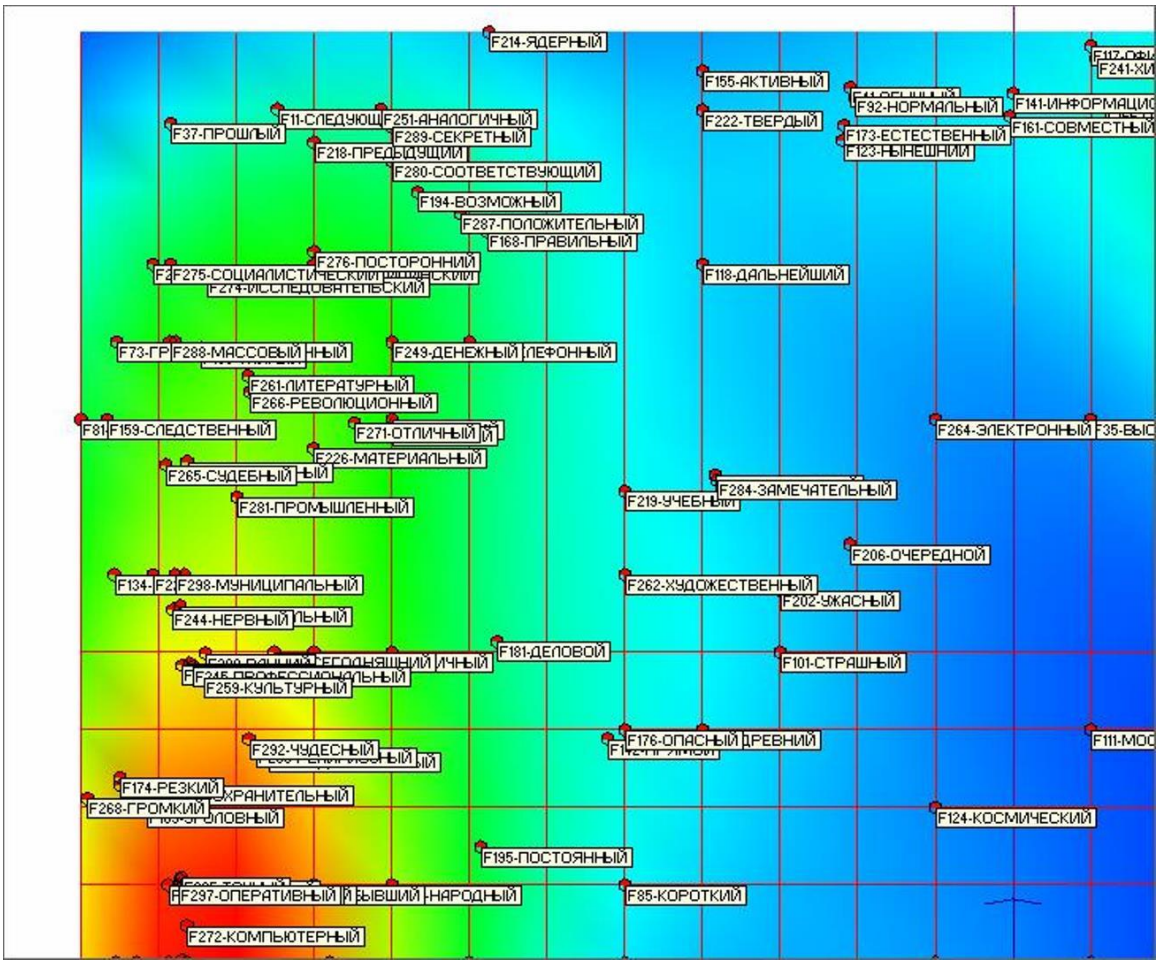

Fig. 11. Extension of the constructed elastic map - close-up - top left corner.

The following figure 12 shows the middle section of the elastic map for the transposed array. Here one can trace the group - ЧУДЕСНЫЙ (WONDERFUL), РЕЛИГИОЗНЫЙ (RELIGIOUS), УДИВИТЕЛЬНЫЙ (AMAZING). Also we can see a group with national and geographic characteristics - КИТАЙСКИЙ (CHINESE), ПОЛЬСКИЙ (POLISH), ЯПОНСКИЙ (JAPANESE).

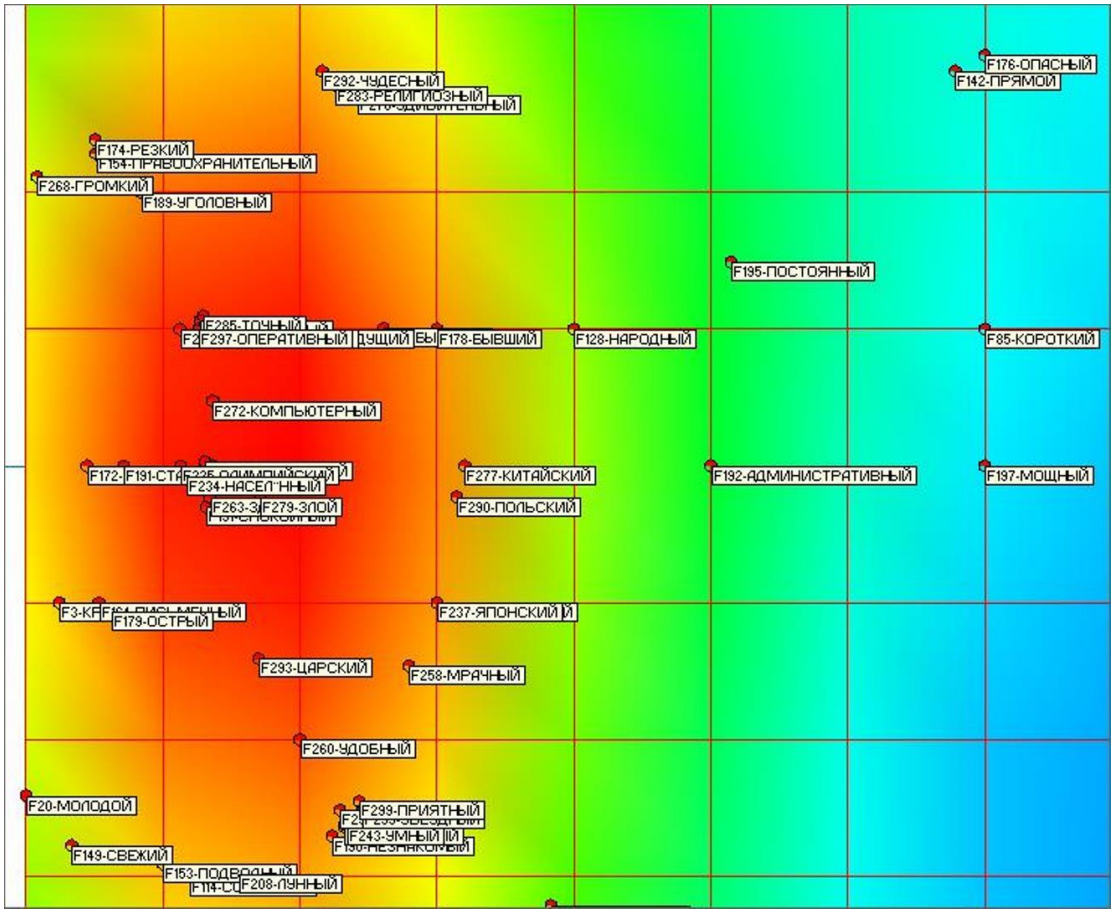

Fig. 12. Extension of the constructed elastic map - close-up – middle of the left edge.

Figure 13 shows the lower part of the left side of the constructed elastic map extension. A number of groups are quite clearly expressed here. A group reflecting the characteristics of the direction - СЕВЕРНЫЙ (NORTH), ЮЖНЫЙ (SOUTH), ЗАПАДНЫЙ (WESTERN), ВОСТОЧНЫЙ (EAST). Group - ГОЛЫЙ (NAKED), ГРЯЗНЫЙ (DIRTY). Group - БЛЕДНЫЙ (PALE), МЯГКИЙ (SOFT), НЕЖНЫЙ (TENDER). The group of adjectives characterizing the temperature is ЛЕДЯНОЙ (ICE), ТЕПЛЫЙ (WARM), ГОРЯЧИЙ (HOT), ХОЛОДНЫЙ (COLD). Group with material characteristics - КАМЕННЫЙ (STONE), ЖЕЛЕЗНЫЙ (IRON).

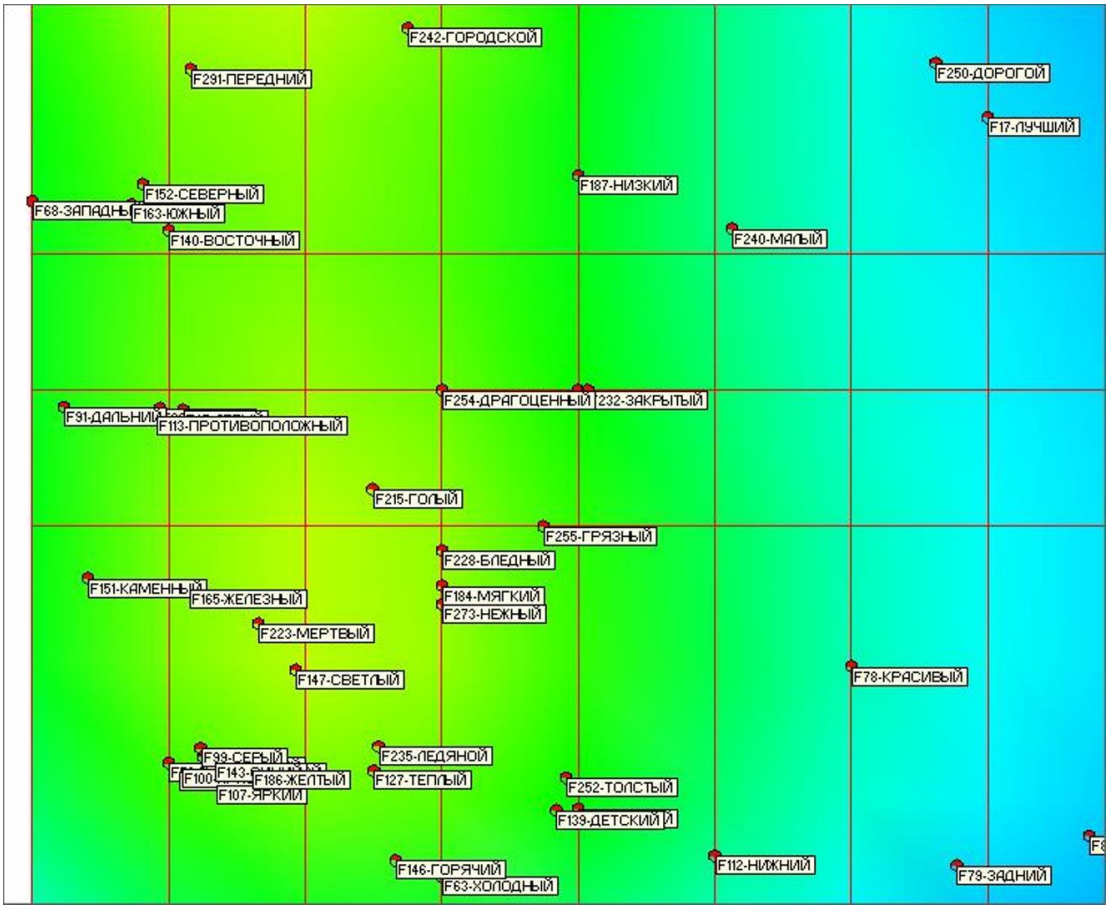

Fig. 13. Extension of the constructed elastic map - close-up - bottom part of the left edge.

Thus, summing up the experiments and the results obtained, it can be argued that the original hypothesis of this study was justified. Recall that we assumed that words that are close in terms of meaning in terms of frequency characteristics should be located close to each other. The implemented approach makes it possible to process volumes of textual information and highlight groups that are similar in semantic characteristics to nouns and adjectives.

4. Conclusion

To analyze the “visual portrait” of a multidimensional data volume, elastic map construction technologies were used. These technologies are methods for mapping points of the initial multidimensional space onto manifolds of smaller dimension embedded in this space. The development of such a map, displayed in the space of the first principal components, allows us to get a "visual portrait" of a multidimensional data volume. Such an image can be organically supplemented by a coloring displaying data density.

This work contains a description of the results of constructing elastic maps for analyzing data volumes consisting of frequencies of joint use of adjectives and nouns. The analysis of the cluster structure for the considered volume of multidimensional data is carried out. A study of the effect of the source data transposition is performed. The initial hypothesis about the proximity in space of signs of words that are close in meaning is confirmed.

The implemented approach allows one to select groups that are similar in semantic characteristics of nouns and adjectives. It should be noted that such an analysis can be useful in the tasks of confronting negative verbal influences such as fake news, hidden propaganda, involvement in sects, verbal manipulation, etc.

References

1. Thomas, J., Cook, K.: Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press, USA (2005).
2. Wong, P., Thomas, J.: Visual Analytics. IEEE Computer Graphics and Applications 24(5), 20-21 (2004).
3. Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F.: Mastering the Information Age – Solving Problems with Visual Analytics. Eurographics Association (2010).
4. Kielman, J., Thomas, J.: Foundations and Frontiers of Visual Analytics. Information Visualization 8(4), 239-314 (2009).
5. Gorban, A. et al.: Principal Manifolds for Data Visualisation and Dimension Reduction. Springer, Berlin – Heidelberg – New York 2007.
6. Gorban, A., Zinovyev, A.: Principal manifolds and graphs in practice: from molecular biology to dynamical systems. International Journal of Neural Systems 20(3), 219–232 (2010).
7. Zinovyev, A.: Visualization of multidimensional data. NGTU, Krasnoyarsk (2000) [in Russian].
8. Zinovyev, A.: Data visualization in political and social sciences, In: Badie, B., Berg-Schlosser, D., Morlino, L. A. (Eds.) INTERNATIONAL ENCYCLOPEDIA OF POLITICAL SCIENCE. SAGE (2011).
9. ViDaExpert, http://bioinfo.curie.fr/projects/vidaexpert, last accessed (01 September 2020).
10. Bondarev, A., Bondarenko, A., Galaktionov, V., Klyshinsky, E.: Visual analysis of clusters for a multidimensional textual dataset. Scientific Visualization 8(3), 1-24 (2016).
11. Bondarev, A., Bondarenko, A., Galaktionov, V.: Visual analysis procedures for multidimensional data. Scientific Visualization 10(4) 109 – 122 (2018). https://doi.org/10.26583/sv.10.4.09
12. Bondarev, A..: The procedures of visual analysis for multidimensional data volumes. ISPRS Archives XLII-2/W12 17-21 (2019). https://doi.org/10.5194/isprs-archives-XLII-2-W12-17-2019
13. Bondarev, A.:Visual analysis and processing of clusters structures in multidimensional datasets. ISPRS Archives XLII-2/W4 151-154 (2017).
14. Bondarev, A., Galaktionov, V.. Applying visual analysis procedures to multidimensional medical data. CEUR Workshop Proceedings 2485 122-126 (2019). https://doi.org/10.30987/graphicon-2019-2-122-126