VISUAL ANALYSIS OF TEXT DATA COLLECTIONS BY FREQUENCIES OF JOINT USE OF WORDS

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

[1] Keldysh Institute of Applied Mathematics RAS, Moscow, Russia - bond@keldysh.ru, vlgal@gin.keldysh.ru
[2]State Res. Institute of Aviation Systems (GosNIIAS), Moscow, Russia - cod@fgosniias.ru

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

ABSTRACT:

The presented research considers 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 words from different parts of speech, for instance “noun + verb” or “adjective + noun”. The volumes of data are 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. 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. Also this approach can be applied to text collections having medical origin.

1. INTRODUCTION

The tasks of analyzing multidimensional data are currently one of the main directions in Computer Science, computational mathematics, mathematical modeling, computer engineering. The huge amount of data that is growing and accumulating in the world requires analysis and processing. Only an analytical study of data, their generalization and identification of key dependencies allows us to see the meaning in their very existence. The need to process, visualize and analyze multidimensional [data has led to the intensive development of visual analytics tools (Wong, Thomas, 2004), (Thomas, Cook, 2005), (Kielman, Thomas, 2009), (Keim et al, 2010). The approaches and methods of visual analytics are constantly evolving and provide users with sufficiently reliable tools for solving many practical problems of multidimensional data exploration. These tasks include the tasks of data classification, cluster detection, identification of key defining parameters, establishing relationships between key parameters, etc.

Visual representation of multidimensional data in a human-readable form is the most visual and effective way to get the maximum amount of information about the data under study. There are a large number of methods for such a visual representation - parallel coordinates, "Chernov's faces", elastic maps, maps of temporal networks, etc.

Visual analytics approaches and methods are usually based on the synthesis of dimensionality reduction algorithms and visual presentation methods. In order to apply the methods of visual representation to the investigated volume of multidimensional data and thus obtain an understanding of the structure and structure of the studied volume of data, it is necessary to map the multidimensional volume of data into the manifolds of lower dimension embedded in the original volume.

Such a mapping can be carried out by building elastic maps (Zinovyev, 2000), (Gorban et al, 2007), (Gorban, Zinovyev, 2010) with different properties of elasticity and their subsequent processing, unfolding and rendering. The elastic maps method is universal; it can be applied to the problems of studying multidimensional data, regardless of the nature of their origin. The creators of the elastic maps approach have found that when mapping the elastic map unfolding into a plane formed by the first two main components, the resulting image reflects the cluster structure of a multidimensional data volume. Thus, a "visual portrait" of the multidimensional data volume is created.

Figure 1 shows an example of a constructed elastic map.

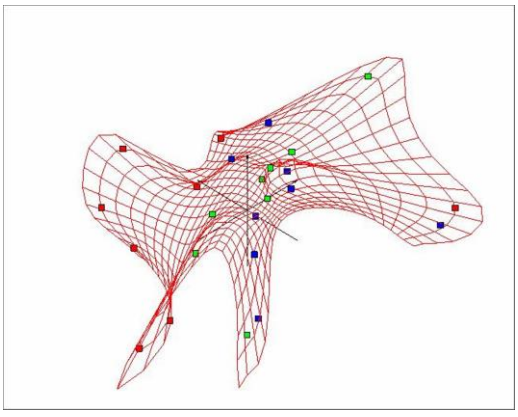

Figure 1. The example of elastic map.

Figure 2 presents an unfolded elastic map in the plane of the first two principal components.

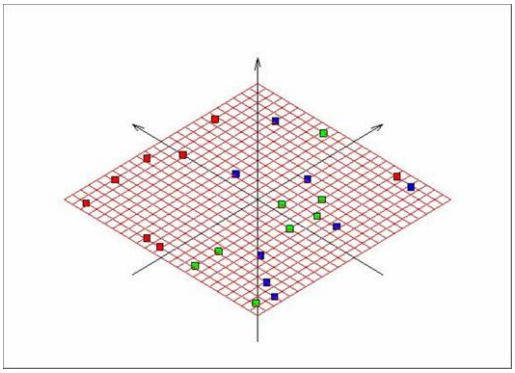

Figure 2. The example of elastic map extension in the space of principal components.

This work is a continuation of research on the development of visual analytics tools for the analysis of multidimensional volumes of numerical and text information. Studies on this topic are presented in (Bondarev et al, 2016), (Bondarev, 2017), (Bondarev, Bondarenko, Galaktionov, 2018), (Bondarev, 2019), (Bondarev, Bondarenko, Galaktionov, 2020), (Bondarev et al, 2020). In the course of research, the construction of elastic maps was tested on a large amount of data of various origins. Among the studied multidimensional data were the characteristics of coal grades, errors of solvers of the open software package OpenFOAM, the results of the analysis of the interaction of supersonic jets, as well as text data volumes. During the research, a number of visual analysis procedures were developed, including procedures such as flotation and quazi-Zoom. The complex application of these procedures makes it possible to improve the results of visual analysis and make the process of obtaining information about the studied volume of multivariate data more efficient.

This work continues a series of works (Bondarev et al, 2016), (Bondarev, Bondarenko, Galaktionov, 2020), (Bondarev et al, 2020) on constructing and transforming elastic maps and conducting experiments with multidimensional data sets, representing the frequencies of the joint use of different parts of speech - adjectives and nouns, verbs and nouns. With the help of certain procedures, text corpora and arrays of shared frequencies are constructed. A visual portrait of the cluster structure of the studied array of multidimensional data was obtained using unfolding and rendering of elastic maps. The study of the influence of the transposition of the initial data has been carried out. The elastic maps technology has shown its efficiency. A study was carried out of a sharp increase in the dimension of the investigated array of frequencies of joint use for adjectives and nouns. It was shown that a sharp increase in the dimension leads to a change in the cluster portrait of the studied data set. The emergence of new subclusters occurs in the cluster structure, the transition of characteristic points from one subcluster to another is observed.

To study the properties of points close to each other on the unfolding of the elastic map, various options for specifying the metric in the space under study were used. Also, various options for determining the centers of clusters formed on the elastic map have been investigated.

An important point should be made. Elastic maps allow you to get an idea of the cluster structure of a multidimensional data cloud without using any clustering algorithms. Clustering algorithms and their settings can introduce additional clutter. In the case of elastic maps, we use only the original data.

2. ELASTIC MAPS

The ideology and algorithms for construction of elastic maps are described in detail (Zinovyev, 2000), (Gorban et al, 2007), (Gorban, Zinovyev, 2010). Elastic map is a system of elastic springs embedded in a multidimensional data space. The method of elastic maps is formulated as an optimization problem, which assumes optimization of a given functional from the relative location of the map and data.

According to (Zinovyev, 2000), 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 elastic properties with respect to stretching and bending. The location of the grid nodes is sought as a result of solving the optimization problem for finding the minimum of the functional consisting of three terms.

The first term is responsible for measure of the proximity of the grid nodes to the data. The second term represents the measure of the stretching of the grid. The third term represents the measure of the curvature of the grid. The last two terms of this functional have coefficients that allow you to adjust the bending and stretching of the elastic map. It is this property that makes it possible to qualitatively change the elastic map, ensuring its maximum approximation to the points of the studied volume of multidimensional data. To represent this in reality, we use the following metaphor. Let's imagine that we can bend and stretch some surface, the properties of which can vary - from hard cardboard to soft paper or cling film. After solving the optimization problem, the constructed elastic map can be unfolded into the plane formed by the first principal components. This way of using elastic maps allows one to obtain a "visual portrait" of the cluster structure of the studied multidimensional volume and is a very effective tool for visual analytics.

The author of the approach (Zinovyev, 2000) has developed the software package (ViDaExpert, 2019), which allows the construction and visual presentation of elastic maps. The main functional features of this software are described in detail in (Zinovyev, 2000). The figures in this article are created by means of this software package.

3. PREPARING TEXT DATA

For numerical experiments, special text collections were created. Pairs from different parts of speech were selected according to the principle "verb + noun" or "noun + adjective". For example, M verbs were selected with the N most related nouns. The data obtained in this way was further considered as a multidimensional data volume, representing M points in N-dimensional space. The numerical values of the resulting matrix were defined as the frequency of sharing.

The selection of data for carrying out numerical experiments for the combinations "verb + noun" was carried out as follows:

At the first stage, text corpora were obtained from news sources. Next, syntactically related pairs of words were extracted from the text corpora. Such information is retrieved as follows:
Step 1 - extract combinations from text. At this step, morphological marking of the text corpus is carried out. Further, verb combinations are selected according to templates. In the formation of this base, nouns take part, unambiguous in terms of speech, but, possibly, ambiguous in case.
Step 2 - compiling a base of word combinations. The combinations obtained at the previous step are brought to normal form, after which their occurrence is calculated. Combinations that occur less than the specified number of times are eliminated from the resulting base.
Step 3 - drawing up a preposition control model. From the combinations obtained in Step 1, those in which the noun is unambiguous in case were selected. For them, the occurrence of pairs of the form "preposition + noun in a given case" was calculated.
Step 4 - getting a control model for verbs. From the base obtained at step 1, we construct a base of combinations of the form "verb + preposition + case of a noun", after which we discard all variants that are prohibited by the preposition model or occur only once.
Step 5 - filtering the collation base. From the base obtained at step 1, we filter out all combinations that do not fit the control model obtained at step 4. At the same time, case ambiguity can be eliminated. As a result, we get the base of the compatibility of verbs with nouns.
The selection of data for combinations of the "noun + adjective" type was carried out in a similar way.

To cut off noise, all combinations with a frequency of occurrence below the specified one are discarded. In addition, only those main words (and their corresponding combinations) are selected for which the cardinality of the set of dependent words exceeds a certain threshold value. This is necessary to filter out noise in the combinations extracted from the collection. The threshold value of the frequency of occurrence allows us to get rid of the combinations that accidentally fell into the base, the number of different combinations guarantees us sufficient statistics for comparisons.

In total, about 7.5 million unique combinations of the form "verb + preposition + noun" and about 2.3 million unique combinations of "noun + adjective" were extracted using regular expressions. For testing, samples of combinations of different parts of speech of different dimensions were made from this set. The numerical values of the resulting matrix are defined as the frequency of sharing. It should be noted that among the selected verbs there were a number of pairs representing similar perfect and imperfective verbs. This was done for additional control due to the assumption that the points corresponding to such pairs should be close to each other in the resulting image.

4. RESULTS OF COMPUTATIONAL EXPERIMENTS

This section presents the history of studies of multidimensional volumes of text data, compiled by the frequencies of the joint occurrence of various parts of speech. Some of the results were previously presented in (Bondarev et al, 2016), (Bondarev, Bondarenko, Galaktionov, 2020), (Bondarev et al, 2020). In these works, the main attention is paid to the study of the possibility of applying the methods of elastic maps for the analysis of thematic proximity of Russian words. The proposed method is based on the analysis of the immediate environment of words. The main hypothesis is 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.

For the initial tests, about 100 verbs were selected with 155 most related nouns. The data obtained in this way is further considered as a multidimensional data volume representing 100 points in 155-dimensional space. The numerical values of the resulting matrix are defined as the frequency of sharing. Nouns correspond to the number of dimensions in a multidimensional volume, and verbs correspond to the number of points in a multidimensional volume.
There were two insidious traps hidden in the data volume under study. The first pitfall, as noted above, was that most verbs were represented by pairs of perfect and imperfective verbs. The goal was to find out if elastic maps are able to recognize this trap and place them on the map in pairs as well. The second trap was the presence in the multidimensional volume of a large lump of "stuck together" words and the study of the possibility of dividing this lump into separate words using an elastic map.

The construction of an elastic map allowed us to successfully overcome the first trap. Figure 3 shows a fragment of an elastic map unfolding, where similar verbs of the perfect and imperfect types are displayed in pairs.

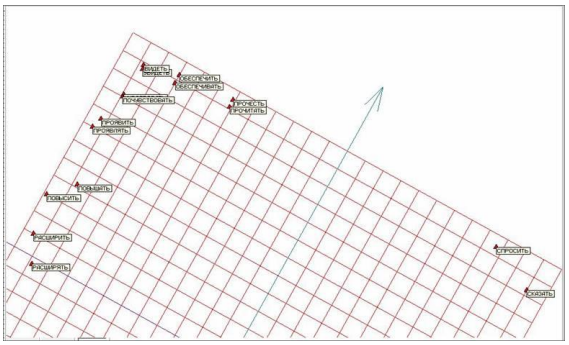

Figure 3. The fragment of elastic map extension with pairs of verbs of perfect and imperfect type

To overcome the second trap, it was necessary to deal with the scalability problem. It should be noted that when building elastic maps in a multidimensional data cloud consisting of condensations and individual distant points, a scalability problem arises. The elastic map will try to adapt to the considered volume as a whole - both to distant points and to areas of concentration, which, of course, cannot work out equally well. Procedures such as flotation and quasi-Zoom have been developed to address this issue and provide a clear understanding of the thickening data.
The flotation procedure is implemented for classification tasks and consists in the fact that the cluster separated from the data cloud is removed from the cloud, and then the procedure for building the map is carried out anew. A similar approach was used in (Niedoba, T., 2014). The approach, called quasi-Zoom, consists in cutting out the area of thickening from the considered multidimensional data cloud and constructing an elastic map for the cropped area. The name is given by the similarity to the procedure used in photography. The application of these procedures allowed us to overcome the second trap. Figure 4 shows a fragment of an elastic map unfolding for an array of frequencies of joint use of 100 verbs and 353 nouns

after two consecutive quasi-Zoom applications. The figure shows that there are practically no “sticky” points left.

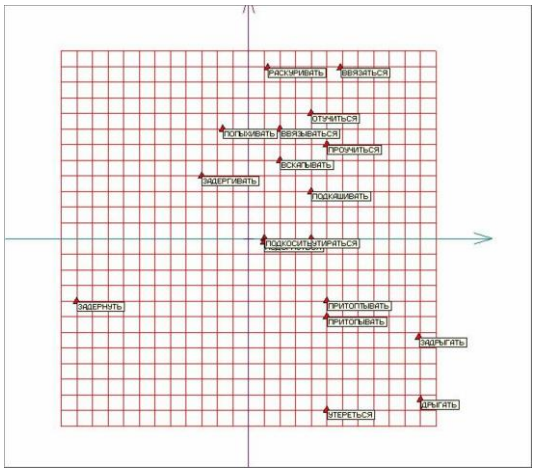

Figure 4. Elastic map extension for data of increased dimension after twice applying the quasi-Zoom procedure.

Transposed arrays were also considered. Here, verbs were already used as dimensions, and nouns were considered as points in multidimensional space. Figure 5 is a general view of a transposed array with annotations.

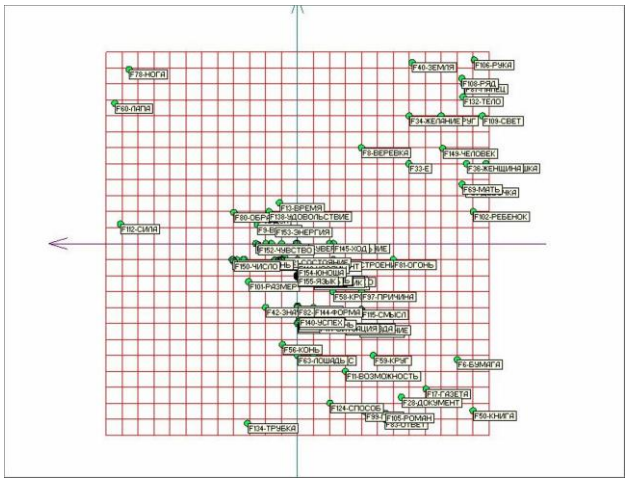

Figure 5. Elastic map extension for transposed array.

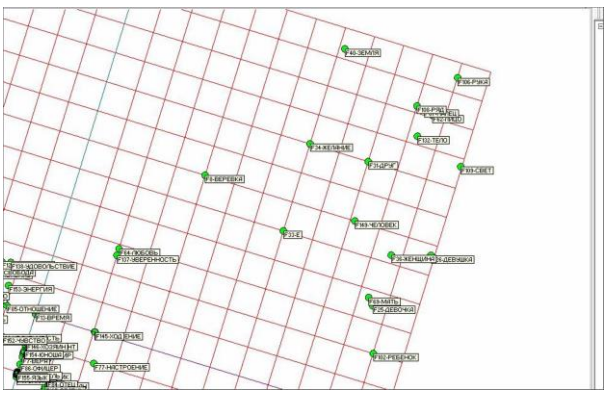

Figure 6. Fragment of elastic map extension with groups of words similar in meaning

No less interesting were data arrays composed of the frequencies of joint use of pairs of parts of speech of the "noun + adjective" type. The number of adjectives was considered as the number of dimensions. The number of nouns was considered as the number of points in a multidimensional space. The frequencies of joint use served as the coordinates of these points in the space thus formed. That is, in this case we are considering 300 points, each of which lies in a 300-dimensional space. When building a sweep of an elastic map, groups of words that are close in meaning appear. Figure 7 shows a fragment of an elastic map sweep.

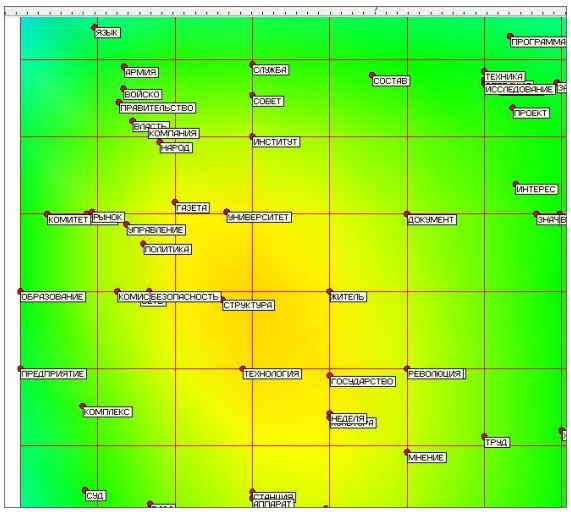

Figure 7. Fragment of elastic map extension “noun + adjective” with groups of words similar in meaning - close-up

The data array under consideration was transposed similarly to the previous example. We studied a transposed data set, where nouns played the role of measurements, and adjectives were considered as points in a multidimensional data set. In this case, we also considered 300 points located in 300-dimensional space. The role of numerical characteristics here was also played by the frequencies of the combined use of adjectives and nouns. Figure 8 shows a close-up of the formation for a transposed array of adjective groups with similar characteristics.

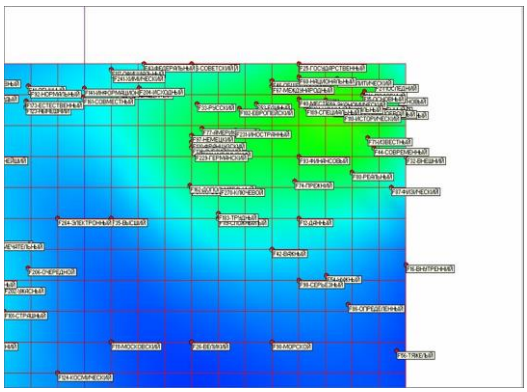

Figure 8. Fragment of elastic map extension for transposed array “noun + adjective” with groups of words similar in meaning - close-up

Similarly to the previous cases, numerical experiments were also carried out here with an increase in the dimension of the studied data set. An array of sampling frequencies of joint use

was built for 2000 adjectives and 1000 nouns. That is, we considered 1000 points lying in 2000-dimensional space. An example of the formed close groups of a noun is shown in Figure 9.

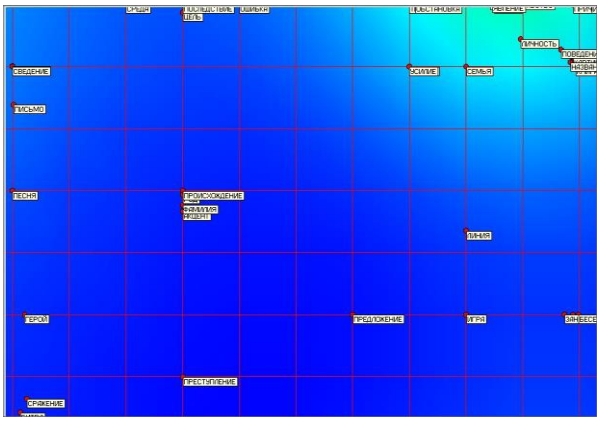

Figure 9. Fragment of elastic map extension for array “noun + adjective” with groups of words similar in meaning - close-up

Here, in the lower right corner, a group is traced – ВСТРЕЧА (MEETING), ЗАНЯТИЕ (LESSON), ЧЕРТА (FEATURE), БЕСЕДА (CONVERSATION), РАЗГОВОР (TALK). In the lower left corner – ГЕРОЙ (HERO), СРАЖЕНИЕ (BATTLE), БИТВА (BATTLE). In the middle on the left – ПРОИСХОЖДЕНИЕ (ORIGIN), РОД (GENUS), ФАМИЛИЯ (SURNAME), АКЦЕНТ (ACCENT).

However, it should be noted that in areas of increased data density, the intersection and overlap of sub-clusters reaches the highest degree. Therefore, it is not possible to identify subclusters without additional procedures.

Normal scaling may not produce the desired result. The presumptive reason can be explained as follows. Note that groups of words sometimes contain words that seem somewhat "alien" in the group. When subclusters are formed from words that are semantically close in terms of frequencies of joint use, the average distance between points in each subcluster is different. This circumstance makes it inevitable that alien points enter the subcluster, which leads to the intersection and overlapping of subclusters in the studied multidimensional data volume.

In such cases, the use of a previously developed system of visual analysis procedures (filtration, flotation, quasi Zoom) may also not give an unambiguously positive result. For a more accurate division of data points in areas of concentration into clusters and subclusters, it is necessary to enter quantitative estimates in order to determine the centers of clusters, determine inter-cluster distances, and determine the average distance between points within a cluster. For further research, it is necessary to introduce the concept of a metric, that is, to specify a method for determining the distance between points of the studied multidimensional data cloud. For these purposes, different metrics are used in multidimensional data volumes. A comparative analysis of various metrics for points lying on the elastic map sweep was carried out according to the "close-far" criterion. That is, the distance between points that are close on the map should be less than between points that are distant from each other. Comparative analysis showed that the best results are provided by the use of the Manhattan metric and the cosine metric. This, in general, was expected, since these metrics are most often used in the analysis of the frequency of occurrence of words. Various ways of determining the center of a cluster were also discussed. As a result, it was found that the best results are obtained by the method of determining the center of the cluster as the arithmetic mean. Figure 10 shows the result of determining the center of the cluster. The center of the cluster is designated in the figure as ЦКЛАСТ. For a more accurate assessment of the proximity of words within clusters and subclusters, quantitative characteristics should be used. The coordinates of cluster centers and average intra-cluster distances can serve as such characteristics. It is also possible to represent subclusters in the form of hyperspheres of different radii. In this case, the radius of the hypersphere, defined, for example, as the maximum distance from the center of the subcluster to its points, will also serve as a defining quantitative characteristic.

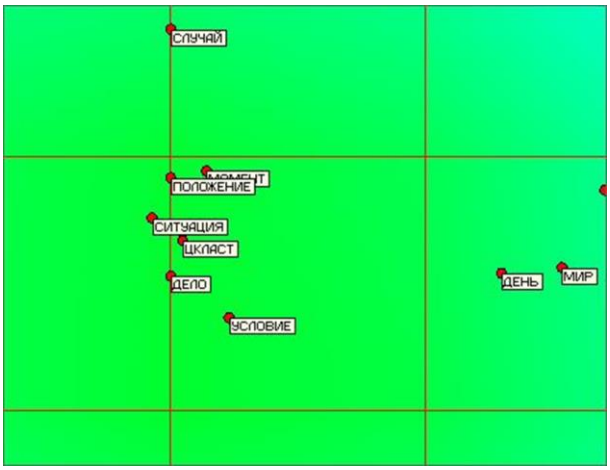

Figure 10. The position of the center of the cluster found by averaging the coordinates - close-up

5. CONCLUSIONS

To analyze the "visual portrait" of a multidimensional volume of data, technologies for constructing elastic maps are used, which are methods of mapping the points of the original multidimensional space onto manifolds of lower dimension embedded in this space. By varying the elastic map surface by successively decreasing the elastic coefficients, it is possible to achieve a better fit of the map adjustment to the multidimensional data cloud. After reducing the bending and stretch coefficients of the elastic map, it becomes softer and more flexible, adapting in the most optimal way to the points of the original multidimensional data volume. The unfolding of such a map, displayed in the space of the first principal components, makes it possible to obtain a "visual portrait" of a multidimensional volume of data. Such an image can be organically complemented by coloring that displays the data density.

The use of technologies for constructing elastic maps for solving cluster analysis problems does not imply any a priori information about the data under study and does not depend on their nature, origin, etc. These properties make it possible to apply technologies for constructing elastic maps to identify cluster structures and proximity of objects when analyzing textual information.

This paper contains a description of the results of constructing elastic maps for analyzing data volumes consisting of frequencies of joint use of various parts of speech - verbs and

nouns, adjectives and nouns. Cases of a sharp increase in the dimension of the considered multidimensional array are considered. Estimates of the distances between near and far points on the elastic map sweep in different metrics are carried out. It was found that the Manhattan metric and the cosine similarity measure show good results. The construction of the center of the cluster of words in different ways was carried out also. It was found that finding the locus of the center of the cluster using arithmetic averaging is fully consistent with the assumed location of the center of the cluster on the scan of the constructed elastic map.

In the course of computational experiments on the study of semantic proximity groups formed on the scan of the elastic map, it was found that in areas of increased data density, where the density of data points is especially large, it is quite difficult to clearly separate subclusters as groups of semantic proximity. Difficulties arise due to the intersection and overlapping of subclusters, as well as different average intracluster distances in different subclusters.

A clear picture of the belonging of an element to a particular cluster or subcluster in terms of semantic proximity can be achieved by applying previously developed data processing and visual analysis procedures (flotation, Quasi-Zoom) in combination with determining the quantitative characteristics of the proximity of elements and the mutual arrangement of clusters and subclusters. A similar approach is expected to be implemented in the future.

Allocation of clusters of words close in the context environment expands the possibilities of contextual search, which can be used in specific tasks of confronting negative verbal influences such as fake news, hidden propaganda, involvement in sects, verbal manipulation, etc. Also this approach can be applied to text collections having medical origin. Currently, in the context of a pandemic, studies of the relationships between medical terms and groups of terms are intensively developing. The above approach may well be applied to such studies. This will require the creation of medical text collections, which is planned for the future.

REFERENCES

Bondarev, A.E. et al, 2016. Visual analysis of clusters for a multidimensional textual dataset. *Scientific Visualization*. 8(3), 1-24.

Bondarev, A.E., 2017. Visual analysis and processing of clusters structures in multidimensional datasets. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W4, 151-154.

Bondarev, A. E., 2019. The procedures of visual analysis for multidimensional data volumes, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W12, 17-21. doi.org/10.5194/isprs-archives-XLII-2-W12-17-2019

Bondarev, A.E., Bondarenko, A.V., Galaktionov, V.A., 2018. Visual analysis procedures for multidimensional data. *Scientific Visualization* 10(4), 109 – 122. doi.org/10.26583/sv.10.4.09.

Bondarev, A.E., Bondarenko, A.V., Galaktionov, V.A., 2020. Visual Analysis of Text Data Volume by Frequencies of Joint Use of Nouns and Adjectives. *Scientific Visualization* 12(4), 9 – 22. doi.org/ 10.26583/sv.12.4.02

Bondarev, A.E., Bondarenko, A.V., Galaktionov, V.A., Shapiro L.Z., 2020. Visual Analysis of Textual Information on the Frequencies of Joint Use of Nouns and Adjectives. *CEUR Workshop Proceedings*, V. 2744, paper20-1 — paper20-10. doi.org/ 10.51130/graphicon-2020-2-3-20

Gorban, A. et al, 2007. *Principal Manifolds for Data Visualisation and Dimension Reduction*, Springer, Berlin – Heidelberg – New York.

Gorban A., Zinovyev A., 2010. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *International Journal of Neural Systems*, 20(3), 219–232.

Keim, D., Kohlhammer, J., Ellis, G., Mansmann, F. 2010 *Mastering the Information Age – Solving Problems with Visual Analytics*. Eurographics Association.

Kielman, J., Thomas, J., 2009. Foundations and Frontiers of Visual Analytics. *Information Visualization* 8(4), 239-314.

Niedoba, T., 2014. Multi-parameter data visualization by means of principal component analysis (PCA) in qualitative evaluation of various coal types. *Physicochemical Problems of Mineral Processing*, 50(2), 575-589.

Thomas, J., Cook, K., 2005. *Illuminating the Path: Research and Development Agenda for Visual Analytics.* IEEE-Press, USA.

ViDaExpert, 2021. bioinfo.curie.fr/projects/vidaexpert (01 March 2021).

Wong, P., Thomas, J., 2004. Visual Analytics. *IEEE Computer Graphics and Applications* 24(5), 20-21.

Zinovyev, A., 2000. *Vizualizacija mnogomernyh dannyh [Visualization of multidimensional data]*. Krasnoyarsk, publ. NGTU. 2000. 180 p. [In Russian]