Scientific Visualization, 2024, volume 16, number 5, pages 75 - 85, DOI: 10.26583/sv.16.5.05

Visualization of Points of a Multidimensional Information Text Array on an Elastic Map for Assessing the Cluster Structure of Data

A.E. Bondarev[1,A]

Keldysh Institute of Applied Mathematics RAS

[1] ORCID: 0000-0003-3681-5212, bond@keldysh.ru

Abstract

The article presents the results of computational experiments on displaying the points of the original multidimensional information array on the elastic map scan to assess the relative positions of semantic proximity areas in order to improve the processing of text information. Elastic maps are considered as a tool for providing analytical work with text information. As previous works show, in order to obtain the required distances corresponding to the cluster picture of the studied multidimensional volume, it is necessary to use the distances on the elastic map, which reflects the cluster portrait of the studied multidimensional data volume. The paper presents the cluster structures of points of the studied multidimensional volume obtained in this way on the elastic map scan in the plane of the first two principal components. An analysis of the relative positions of clusters of different configurations at different points in time is presented.

Keywords: Multidimensional text data, cluster structure, elastic maps, cluster position analysis.

1. Introduction

This paper presents the results of computational experiments on cluster analysis for multidimensional information arrays. Multidimensional information arrays represent text information, being in digital form the frequencies of joint use of words from different parts of speech (for example, noun + adjective). To obtain such a multidimensional array, an analysis of text collections is carried out. As a result, for a 300-dimensional array constructed in this way, we consider 300 points in a space of 300 dimensions. This work is a continuation of the works [19-26] and largely uses algorithms, results and developments from previous works.

At the present stage, the study and analysis of multidimensional data volumes is becoming an extremely urgent task. Analysis of multidimensional data has been a pressing problem for quite a long time. Methods of data analysis and visual analytics were developed for such studies. The use of these methods makes it possible to practically learn the structure of the studied volume of multidimensional data, see its cluster picture, determine areas of data condensation (clusters), etc.

The study of multidimensional data has formed a new interdisciplinary field of research known as visual analytics (Visual Analytics). Visual analytics is a set of methods and approaches aimed at analyzing and visually representing multidimensional data regardless of the nature of their origin. The main concepts of visual analytics are presented in the first works on this discipline [1-4]. Approaches and methods of visual analytics allow solving a number of practical problems of studying multidimensional data. Such problems include data classification, cluster detection, identification of key defining parameters, establishment of relationships between key parameters, etc.

To display a multidimensional volume of data into lower-dimensional manifolds embedded in the original volume, approaches of self-organizing maps have been developed [5-7]. Such maps also include the so-called elastic maps (Elastic Maps). The theory of elastic maps was developed by A. Gorban and A. Zinoviev together with colleagues and is described in sufficient detail in works [8-14]. Elastic maps have found quite wide application in practical analysis of multidimensional data, and in various fields - analysis of economic data, analysis of minerals, analysis of medical data, etc. Examples of studies are presented in works [15-18]. The most important property of elastic maps is the ability to successfully use them for any multidimensional data regardless of the nature of their origin. Elastic maps acquire particular efficiency in solving problems of cluster analysis of multidimensional data when used together with the principal component analysis (PCA). Displaying an elastic map and its scan in the space formed by the first three principal components in clustering and classification problems allows more accurately and clearly determining the cluster structure of the studied multidimensional data volumes. Figure 1 shows an example of an elastic map unfolding onto a plane located in the space of principal components with coloring by data density. Such a representation allows one to determine the cluster structure of the studied volume of multidimensional data without using special clustering algorithms.

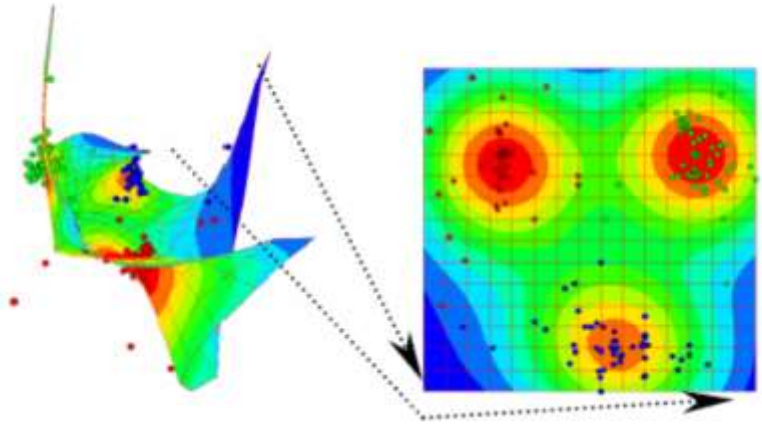

Fig. 1. An example of an elastic map and its development onto a plane with coloring by data density.

Of great importance is the application of elastic map construction to the problems of multidimensional text data analysis. This direction appeared from the need to solve practical problems of cluster analysis and is becoming increasingly relevant. The application of elastic map construction to the problems of multidimensional text data analysis was first implemented in practice by A.E. Bondarev and V.A. Galaktionov at the Keldysh Institute of Applied Mathematics of the Russian Academy of Sciences. Arrays of frequencies of joint use of various parts of speech obtained from text collections were used as the studied volumes of multidimensional data. The results of the studies are presented in detail in [19-26]. However, in the process of solving the problems of cluster analysis of multidimensional information volumes, one extremely important circumstance was revealed. The fact is that the picture of cluster distribution in a multidimensional volume of text data obtained in numerical experiments exists only on the expansion of elastic maps in the space of the first two principal components. Determination of specific intracluster and intercluster distances should be carried out on digital data. Such data are the coordinates of the points of a multidimensional volume in a multidimensional space. However, as shown by the works [24-26], the determination of intracluster and intercluster distances based on the initial coordinates of the points of a multidimensional space does not give reliable results. To obtain the required distances, it is necessary to use the distances on an elastic map, which reflects the cluster portrait of the studied multidimensional data volume. This work is devoted to this direction of research.

2. Elastic maps and application of the principal component analysis (PCA) in visual analysis of multidimensional text information arrays

Elastic maps are a logical development of Kohonen maps. The ideology and algorithms for constructing elastic maps are presented in detail in [8 - 14]. Such a map is a system of elastic springs embedded in a multidimensional data space. The elastic map method is formulated as an optimization problem that involves optimizing a given functionality from the mutual arrangement of the map and data.

According to [8 - 14], the basis for constructing an elastic map is a two-dimensional rectangular grid G embedded in a multidimensional space, which 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 an optimization problem to find the minimum of the functional:

$$D = \frac{D_1}{|X|} + \lambda \frac{D_2}{m} + \mu \frac{D_3}{m} \to min,$$

where $|X|$ is the number of points in the multidimensional data volume X; m is the number of grid nodes, λ, μ are the elasticity coefficients responsible for the stretching and curvature of the grid, respectively; D_1, D_2, D_3 are the terms responsible for the properties of the grid.

$$D_1 = \sum_{ij} \sum_{x \in K_{ij}} \left|\left| x - r^{ij} \right|\right|^2 .$$

D_1 is a measure of the proximity of the grid nodes to the data.

Here Kij are the subsets of points from X for which the grid node r_{ij} is the closest:

$$x \overset{\Pi}{\to} r^{ij}, \qquad \left|\left| x - r^{ij} \right|\right|^2 \to \min, \qquad K_{ij} = \left\{ x \in X, x \overset{\Pi}{\to} r^{ij} \right\},$$

The term D_2 represents the measure of the grid stretch:

$$D_2 = \sum_{ij} \left\| r^{ij} - r^{i,j+1} \right\|^2 + \sum_{ij} \left\| r^{ij} - r^{i+1,j} \right\|^2 .$$

The term D_3 represents the measure of the curvature of the mesh:

$$D_3 = \sum_{ij} \left\| 2r^{ij} - r^{i,j-1} - r^{i,j+1} \right\|^2 + + \sum_{ij} \left\| 2r^{ij} - r^{i-1,j} - r^{i+1,j} \right\|^2 .$$

Varying the elasticity parameters involves constructing elastic maps with a sequential decrease in elasticity coefficients, due to which the map becomes softer and more flexible, most optimally adapting to the points of the original multidimensional data volume. After construction, the elastic map can be unfolded into a plane to observe the cluster structure in the studied data volume. On the unfolded plane, the distribution of data density by the elastic map can be displayed by coloring. In some cases, such coloring can be very useful. Elastic maps are especially effective when used together with the principal component analysis (PCA). Displaying the elastic map and its unfolding in the space formed by the first three principal components allows for a sharp improvement in the results, especially in clustering and classification problems. The use of elastic maps allows for a more accurate and precise determination of the cluster structure of the studied multidimensional data volumes.

The described approach was applied many times to multidimensional text data arrays. Multidimensional data obtained from text collections and representing frequencies of joint use of different parts of speech were considered. For example, 300 nouns and 300 adjectives. Adjectives were considered as dimensions, and nouns as points in the space of dimensions. That is, 300 points in a 300-dimensional space were considered. Figure 2 shows a typical view of the elastic map.

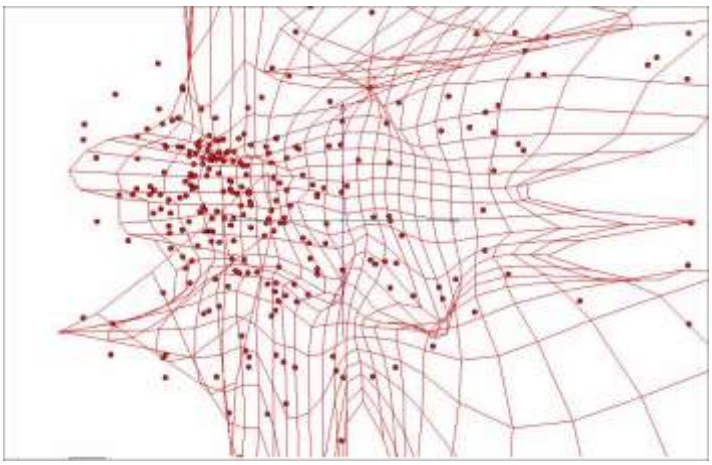

Fig. 2. Typical appearance of an elastic frequency map of joint use.

3. Construction of the initial data set and testing of the cluster arrangement

Previously, in previous works [20-26] it was shown that the approach of estimating intercluster distances by constructing hyperspheres in the space of the initial data does not provide the desired result. All numerical studies were conducted using two types of metrics that define the distance between objects - the Manhattan metric and the cosine metric. To estimate intercluster distances, it is necessary to construct other methods of estimation. Based on the results of works [20-26], it was concluded that the problems with studying cluster properties are associated with the fact that the initial qualitative information about clusters is taken from the elastic map scan, and the initial coordinates of points in multidimensional space are used for calculations. It was decided to use the coordinates of the annotated points on the elastic map scan in the future due to the fact that they directly reflect the cluster picture of the studied data volume.

The real data were obtained from text corpora of news information. To obtain information from text corpora, the procedures described in detail in [20] were used. To obtain the required multidimensional arrays, groups of combinations of "noun + adjective" were used. 300 nouns and 300 adjectives were selected. The frequencies of joint use of a noun and an adjective in each studied multidimensional array were used as a digital value. Thus, as in previous experiments, we were able to consider adjectives as coordinate dimensions in a multidimensional space, and nouns as points in this space. 300 points located in a 300-dimensional space were considered. In this way, 5 multidimensional arrays were constructed - for March 2005, for April 2005, for May 2005, for May 2006 and for May 2007. The first three arrays make it possible to track the evolution of the cluster structure of the multidimensional volume after a month. The third, fourth and fifth arrays allow us to trace the evolution of the cluster structure of the multidimensional volume over a year. For all arrays, elastic maps and their scans in the space of the first principal components were constructed.

4. Results of computational experiments

The computational experiments included the construction of elastic maps and their developments at different moments in time, the selection of clusters, finding their centers and radii. Let us recall that in this problem the Euclidean metric was used to determine distances, the center of each specific cluster was determined as the arithmetic mean of the points included in this cluster. The distance from the center of the cluster to the most remote point of the cluster was chosen as the characteristic size of the cluster.

Let us begin our consideration of the results of the computational experiments with the elastic map sweep for the time moment 2005-4 (April 2005). Here in the multidimensional data array on the lower right edge of the elastic map sweep there are two clearly defined clus-

ters, shown in Fig. 3, where a fragment of the sweep containing these clusters together with their calculated centers is shown in close-up.

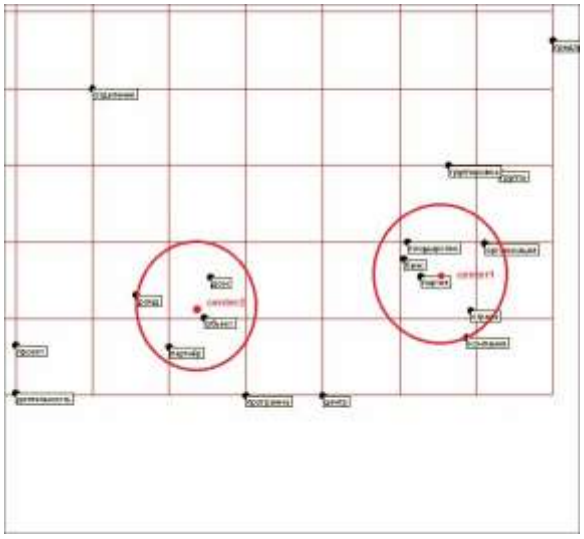

Fig. 3. Lower right edge clusters for the studied multidimensional array at time 2005-4 with cluster centers.

Figure 4 shows a fragment of the elastic map scan with annotations, corresponding to the time point 2005-5 (May 2005) with two clusters along the lower right edge and cluster centers.

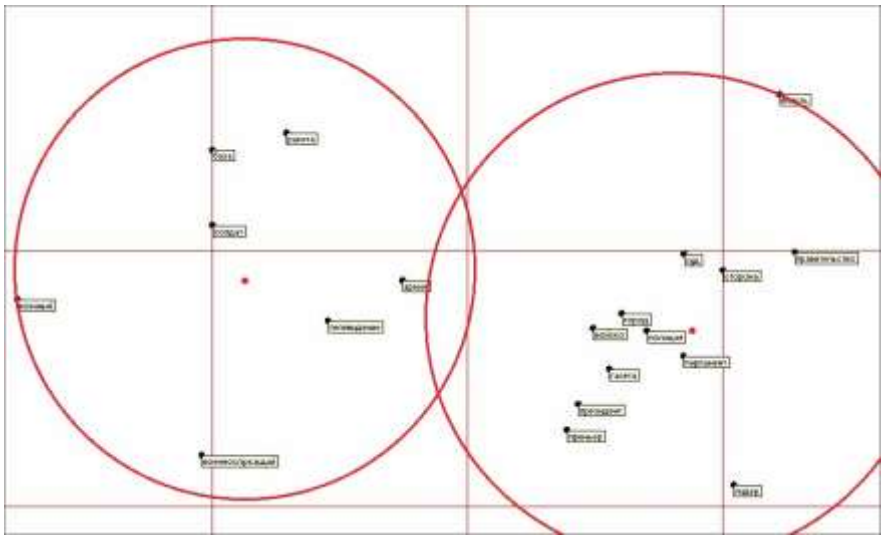

Fig. 4. Clusters of the upper right edge for the studied multidimensional array at the time 2005-05 with cluster centers and radii.

The most distant points after calculating the distances from the cluster center are the points "AUTHORITY" and "MILITARY". According to the results obtained, in this case a small area of cluster intersection is formed, but not a single point of the studied array of multidimensional data falls into this point.

Now let's look at the results for the time point 2006-05. Figure 5 shows the elastic map unfolding with annotations corresponding to the time point 2006-5 (May 2006).

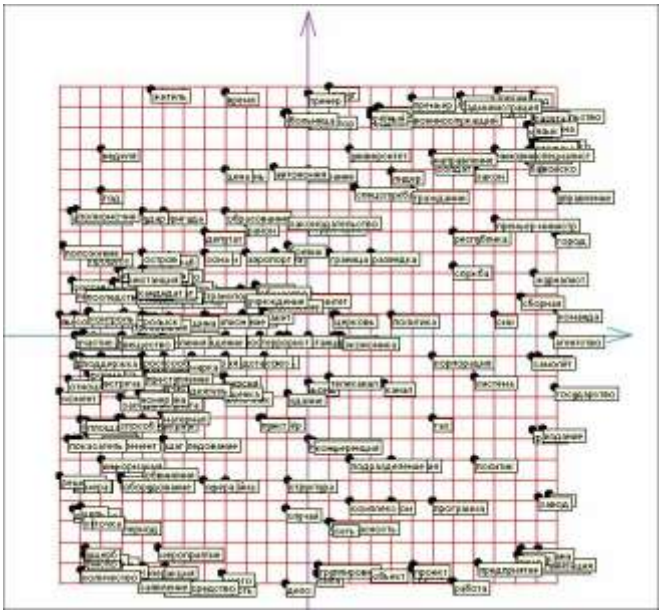

Fig.5. Annotated expansion of the elastic map at the time 2006-05.

Here in the multidimensional data array on the right edge in The upper and lower parts of the elastic map scan have two clearly defined clusters. The clusters of the upper part are shown in Figure 6.

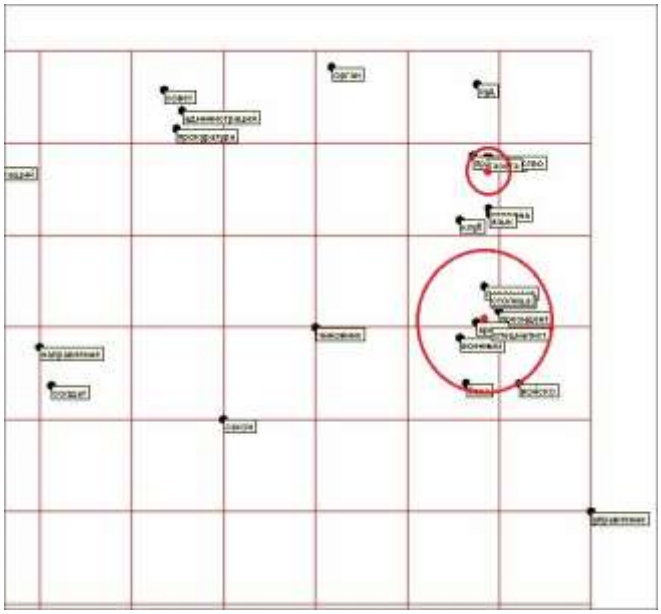

Fig. 6. Clusters of the upper right edge for the studied multidimensional array at the time point 2006-05 with cluster centers.

Figure 7 shows the lower right clusters at time 2006-5 (May 2006).

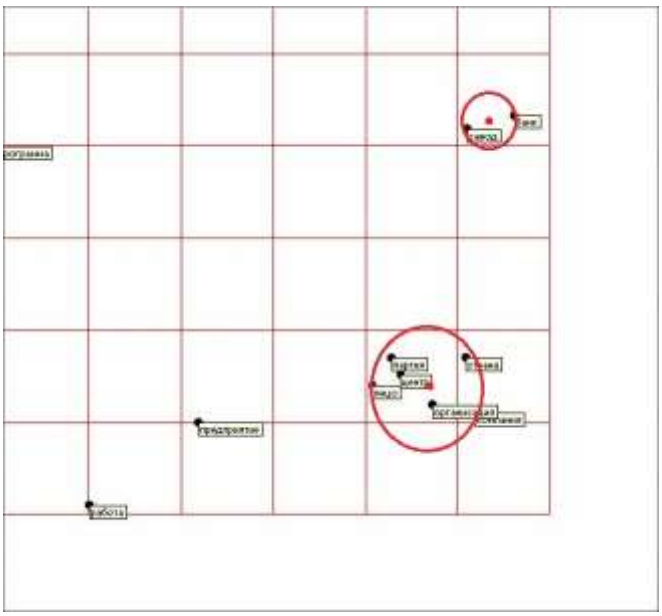

Fig. 7. Lower right edge clusters for the studied multidimensional array at the time point 2006-05 with cluster centers.

The presented results show that in the considered cases there is no intersection of clusters and the emergence of “controversial” points.

Let us consider the cluster picture at the time point 2007-05. The annotated expansion of the elastic map is presented in Figure 8.

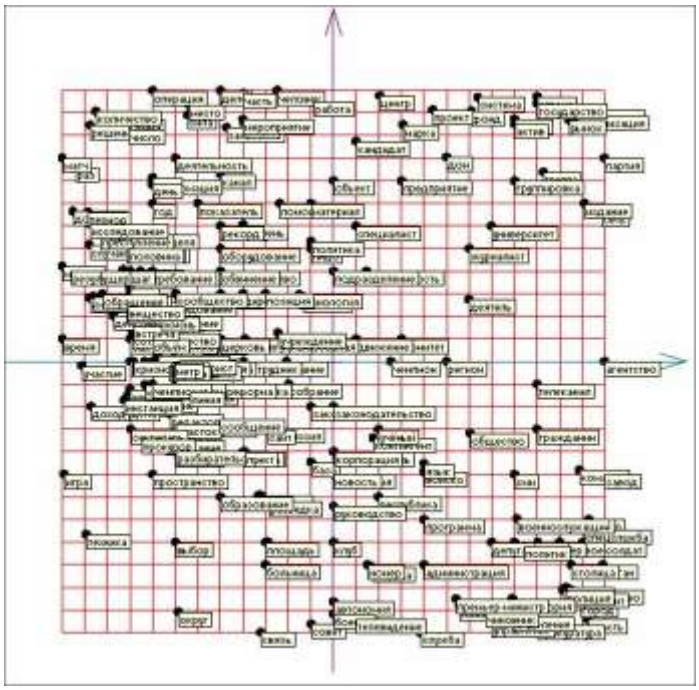

Fig. 8. Annotated expansion of the elastic map at the time 2007-05.

Three clusters are formed in the upper part of the right corner of the elastic map development. Their appearance is shown in Figure 9.

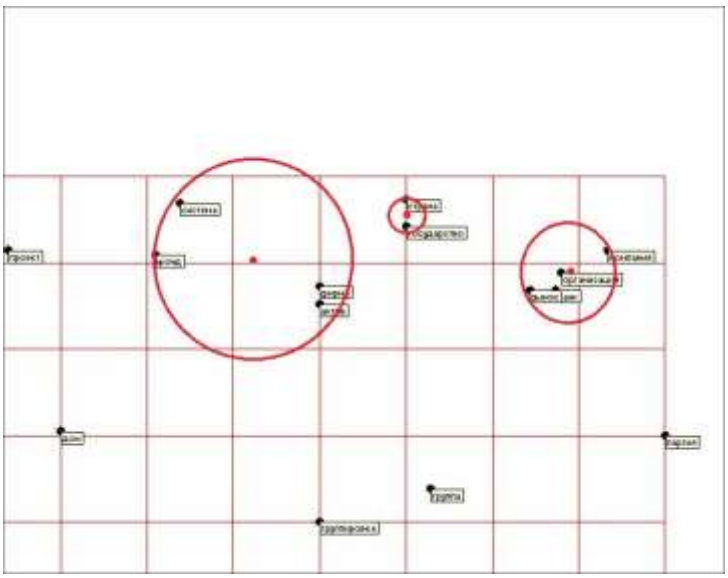

Fig. 9. Clusters of the upper part of the right corner of the elastic map scan at the time of 2007-05 in the form of circles and their centers.

Now let's imagine the lower part of the right corner of the elastic map scan for the same time 2007-05. Here the picture is presented in Figure 10. It should be noted that in this case the clusters have a more elongated shape than in the previous cases.

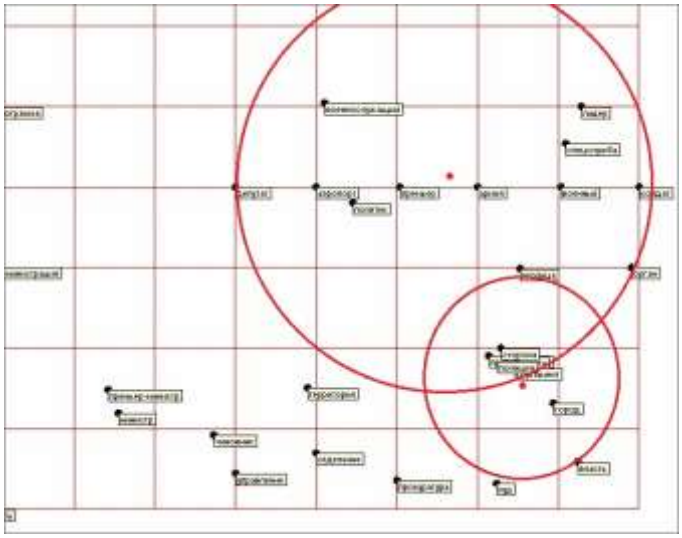

Fig.10 . Clusters of the lower part of the right corner of the elastic map scan at the time of 2007-05 in the form of circles and their centers.

It is here that we encounter a situation where clusters constructed according to the rules we have defined intersect. In this intersection, the larger cluster includes the "extra" points "CAPITAL" and "ORGAN", as well as the points of the second smaller cluster. The result obtained shows that the rules for constructing clusters in such cases must be adjusted so that they can take into account "stretched" clusters that are far from the shape of circles.

5. Conclusion

Elastic maps, when presented in a sweep onto a plane formed by the first two principal components, allow one to see the cluster picture of the studied multidimensional volume. An array of frequencies of joint use of different parts of speech (for example, "adjective + noun") obtained from text collections is used as a multidimensional volume. To analyze the cluster

structure, the coordinates of the points of the multidimensional volume obtained from the sweep of the elastic map are used. The mutual arrangements of clusters arising at different moments in time are considered. Clusters are presented as circles. To construct the radii, the principle of choosing the distance from the center of the cluster to its most distant point was used.

As a result of the computational experiments carried out, it can be stated that:

- the use of elastic maps and their scans is an effective tool for an analyst when studying multidimensional text information;
- the procedures for constructing clusters and their mutual arrangement must be improved in such a way that they can reflect “extended” clusters, problems of cluster intersection, and problems of cluster nesting;
- it is necessary to develop, on the basis of these procedures, a clear methodology for the application of elastic maps and their scans to the analysis of the cluster structure of multidimensional information data.

References

1. Thomas J. and Cook K. 2005 Illuminating the Path: Research and Development Agenda for Visual Analytics (IEEE-Press)
2. Wong PC, Thomas J. Visual Analytics // IEEE Computer Graphics and Applications. 2004. V. 24, N. 5. - P. 20-21.
3. Keim D., Kohlhammer J., Ellis G. and Mansmann F. (Eds.) Mastering the Information Age – Solving Problems with Visual Analytics, Eurographics Association, 2010.
4. Kielman, J. and Thomas, J. (Guest Eds.) (2009). Special Issue: Foundations and Frontiers of Visual Analytics / Information Visualization, Volume 8, Number 4, p. 239-314.
5. T. Kohonen, Self-Organizing Maps (Third Extended Edition), New York, 2001, 501 pages.
6. Debock G., Kohonen T. Analysis of financial data using self-organizing maps, Alpina Publisher, 2001, 317 p.
7. Kohonen T. Self-organizing maps. - M.: BINOM. Laboratory of knowledge, 2008. - 655 p.
8. Gorban A. et al. Principal Manifolds for Data Visualization and Dimension Reduction, LNCSE 58, Springer, Berlin – Heidelberg – New York, 2007.
9. AN Gorban, AY Zinovyev, Principal Graphs and Manifolds, From: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques, Olivas ES et al Eds. Information Science Reference, IGI Global: Hershey, PA, USA, 2009. 28-59.
10. Zinovyev A. Vizualizacija multidimensional dannyh [Visualization of multidimensional data]. Krasnoyarsk, publ. N.G.T.U. 2000. 180 p. [In Russian}
11. Zinovyev A. Data visualization in political and social sciences, In: SAGE “International Encyclopedia of Political Science”, Badie, B., Berg-Schlosser, D., Morlino, LA (Eds.), 2011.
12. Pitenko A.A. Neural network analysis in geoinformation systems. Krasnoyarsk, Publ. KSTU, 2000. 97 p.
13. Rossiev A.A. Iterative modeling of incomplete data using low-dimensional manifolds. Krasnoyarsk, KSTU Publ., 2000. 83 s.
14. ViDaExpert, http://bioinfo.curie.fr/projects/vidaexpert, last accessed (01 March 2020).
15. Niedoba T., Multi-parameter data visualization by means of principal component analysis (PCA) in qualitative evaluation of various coal types, Physicochemical Problems of Mineral Processing, vol. 50, iss. 2, pp. 575-589, 2014.
16. H. Shaban, S. Tavoularis, Identification of flow regime in vertical upward air–water pipe flow using differential pressure signals and elastic maps, International Journal of Multiphase Flow 61 (2014) 62-72.

17. H. Shaban, S. Tavoularis , Measurement of gas and liquid flow rates in two-phase pipe flows by the application of machine learning techniques to differential pressure signals, International Journal of Multiphase Flow 67(2014), 106-117

18. M. Resta, Computational Intelligence Paradigms in Economic and Financial Decision Making, Series Intelligent Systems Reference Library, Volume 99, Springer International Publishing, Switzerland 2016

19. Bondarev AE, Bondarenko AV, Galaktionov VA, Klyshinsky ES Visual analysis of clusters for a multidimensional textual dataset / Scientific Visualization. V.8, No. 3, pp.1-24, 2016, URL: http://sv-journal.org/2016-3/index.php?lang=en

20. AE Bondarev, AV Bondarenko, VA Galaktionov (2018) Visual analysis procedures for multidimensional data. Scientific Visualization 10.4: 109 - 122, DOI: 10.26583/ sv.10.4.09 http://www.sv-journal.org/2018-4/09?lang=en

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

22. Bondarev AE Visual analysis and processing of clusters structures in multidimensional datasets // Proceedings of the 2nd International ISPRS Workshop on PSBB, 15–17 May 2017, Moscow, Russia, ISPRS Archives, Volume XLII-2/W4, 2017, pp.151-154. http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W4/151/2017/

23. Bondarev A.E., Galaktionov V.A., Shapiro L.Z. Processing and visual analysis of multidimensional data / Scientific visualization, Vol.9, No.5, 2017, pp. 86-104. http :// sv - journal . org /2017-5/08/ index . php ? lang = ru

24. Alexander Bondarev, Alexander Bondarenko, Vladimir Galaktionov , Lev Shapiro. Visual Analysis of Textual Information on the Frequencies of Joint Use of Nouns and Adjectives // CEUR Workshop Proceedings, V. 2744, Proc. of the 30th International Conference on Computer Graphics and Machine Vision GraphiCon 2020, Saint Petersburg, Russia, September 22-25, 2020, p. paper20-1 - paper20-10, DOI: 10.51130/graphicon-2020-2-3-20

25. AE Bondarev, AV Bondarenko, VA Galaktionov. Visual Analysis of Text Data Volume by Frequencies of Joint Use of Nouns and Adjectives (2020). Scientific Visualization 12.4: 9 - 22, DOI: 10.26583/sv.12.4.02

26. Bondarev, AE, Bondarenko, AV, and Galaktionov, VA: Visual analysis of text data collections by frequencies of joint use of words, Int. Arch. Photogramm . Remote Sens. Spatial Inf. Sci., XLIV-2/W1-2021, 21–26, 2021. https://doi.org/10.5194/isprs-archives-XLIV-2-W1-2021-21-2021