VISUAL ANALYSIS OF TIME-VARYING MULTIDIMENSIONAL DATA SETS

A.E. Bondarev

Keldysh Institute of Applied Mathematics RAS, Moscow, Russia – bond@keldysh.ru

Commission II, WG II/8

KEY WORDS: Multidimensional Data, Time-Varying Data, Elastic Maps, Frequencies of Joint Use, Cluster Structures.

ABSTRACT:

This paper contains a description of computational experiments on the application of elastic maps to the analysis of time-varying volumes of textual information. Elastic maps are considered as a tool to provide analytical work with textual information and large information arrays of data. 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”. We consider text collections in Russian for experiments. Previously, static information arrays were mostly considered. It is for them methods of data analysis and methods of visual analytics were developed. Nevertheless, data comes in all the time in various areas of human activity. And in practice it is necessary to know how the cluster picture of multidimensional data volume changes over time. The paper describes the numerical experiments for real time-varying multidimensional data sets. Such experiments allows to analyze the evolution of cluster structure for multidimensional data and to trace the evolution for separate cluster.

1. INTRODUCTION

This paper contains a description of computational experiments on the application of elastic maps to the analysis of time-varying volumes of textual information. Elastic maps are considered as a tool to provide analytical work with textual information and large information arrays of data.
Consider why elastic maps are such an effective tool for analyzing multivariate data. This is primarily because elastic maps allow us to obtain a "cluster portrait" of multivariate data and to study their internal structure regardless of the nature of the data origin. Thus, we can use elastic maps for a very wide range of tasks of multidimensional data analysis. Among the main applications of elastic maps are the analysis of multivariate data in areas such as:

- environmental monitoring;
- content of text collections;
- medicine and prognostic diagnostics;
- analysis of the main influencing factors on the performance of any complex technics.

Elastic maps are particularly important for the visual analysis of changes in multidimensional data over time. This is the phenomenon addressed in this paper, where for the first time elastic maps are applied to time-varying multivariate data sets. This makes it possible to construct elastic map sets for any monitoring task, including environmental monitoring.

At the present stage, the study and analysis of dynamic, i.e. changing over time, multidimensional volumes of data becomes an extremely urgent task. Analysis of multidimensional data has been a topical problem for quite a long time. But, previously, static information arrays were mainly considered. It is for them methods of data analysis and methods of visual analytics were developed. Nevertheless, data comes in all the time in various areas of human activity. And in practice it is necessary to know how the cluster picture of multidimensional data volume changes over time.

The need to process, visualize and analyze multidimensional data has led to the intensive development of visual analytics tools (Thomas, Cook, 2005), (Kielman, Thomas, 2009), (Keim et al, 2010). Visual analytics approaches and methods are constantly evolving and provide users with a sufficiently robust means to solve many practical multidimensional data research problems. Such tasks can include data classification tasks, cluster detection, identification of key defining parameters, and establishment of relationships between key parameters. Visual analytics is an area of interdisciplinary research, as it allows you to study data regardless of the nature of their origin. Visual analytics approaches are a synthesis of several algorithms of dimensionality reduction and visual representation of multidimensional data in the nested in the original volume manifolds of lower dimensionality.

An important tool for visual analytics was the emergence of elastic maps. The theory of elastic maps was proposed by a group of authors in the early 2000s. The basics of the theory were outlined in (Zinovyev, 2000), (Gorban et al., 2007), (Gorban and Zinovyev, 2010). The construction of elastic maps with different elasticity or elasticity properties and their subsequent processing, unwrapping and rendering allows one to obtain information on the cluster structure of the multidimensional data under study. The method of elastic maps is universal; it can be applied to the problems of studying multidimensional data regardless of the nature of their origin. A special place in the tasks of elastic maps is occupied by the tasks of analysis of multidimensional textual data. This direction is becoming more and more relevant. Such an application was first put into practice in (Bondarev et al., 2016), (Bondarev, 2017), (Bondarev et al., 2018), (Bondarev et al., 2020), (Bondarev et al., 2021). The volumes of multivariate data under study were arrays of frequencies of joint use of different parts of speech obtained from textual collections. However, in all of the aforementioned works, elastic maps were constructed in order to analyze static arrays exclusively. That is,

in the studies, the object of study was the arrays that do not change over time. However, most important in the tasks of multivariate data analysis is the study of dynamic arrays, i.e., arrays that change over time. Indeed, the analysis of multidimensional data of any nature is primarily aimed not only at studying the cluster pattern of a multidimensional data volume, but also at changing this cluster pattern. The study of such changes over time is able to provide indications of the factors that have the greatest influence on the cluster picture and ultimately can lead in practice to the possibility of controlling processes. The purpose of this work is to develop and practically test an approach of using elastic maps to analyze real dynamic volumes of multidimensional textual information data. The developed computational methods and procedures suggest that it would be most reasonable to consider multidimensional data changing over time as a series of "snapshots" taken at different points in time. A sketch of such a representation is shown in Figure 1.

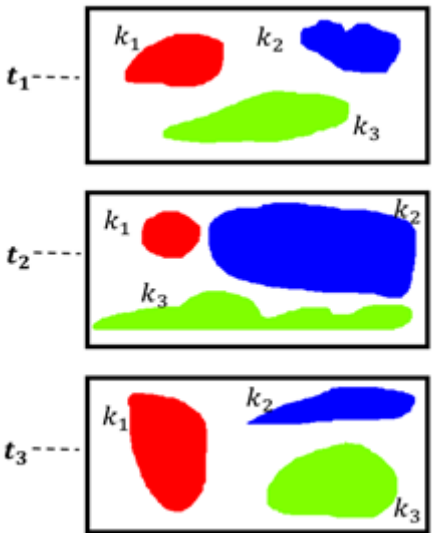

Figure 1. Schematic of the application of a series of snapshots derived from extensions of elastic maps for the visual analysis of dynamically changing data.

This approach was tested earlier on previously prepared synthetic (artificial) data. An initial array of nouns and adjectives was selected and 4 more arrays were constructed to simulate the changes in the multidimensional volume of data under study over time.

Elastic maps and their extensions in the space of first principal components were calculated for all constructed sets of artificial data. We analyzed changes in the cluster picture of the studied multidimensional data volume. The changes occurring to individual clusters were considered. According to the results of numerical experiments, it can be argued that this approach of visual analysis is workable and can serve as a working tool for studying dynamic arrays.

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.

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, 2023), 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. ANALYSIS OF THE EVOLUTION OF THE CLUSTER PICTURE

Let us consider the results of computational experiments to analyze the evolution of the cluster picture on real dynamic information arrays. Real dynamic information arrays were used for computational experiments. The general computational approach remained the same as previously used for synthetic (artificial) data. The real data were obtained from text corpuses of news information. The procedures described in detail in (Bondarev et al., 2016) were used to obtain information from text corpuses. Groups of noun + adjective combinations were used to obtain the necessary multivariate arrays. 300 nouns and 300 adjectives were selected. The frequencies of noun and adjective combinations in each multivariate array under study were used as the numerical value. Thus, like in previous experiments, we were able to consider adjectives as coordinate dimensions in multidimensional space, and nouns as points in this space. We considered 300 points in 300-dimensional space. It should be noted that when choosing nouns, preference was given to nouns from the political-economic block. A total of 5 multidimensional arrays were thus constructed - for March 2005, for April 2005, for May 2005, for May 2006, and for May 2007. The arrays constructed with such temporal choice make it possible to analyze different time intervals. The first three arrays allow us to trace the evolution of the cluster structure of the multivariate volume through the month. The third, fourth and fifth arrays allow us to trace the evolution of the cluster structure of the multidimensional volume through one year. For all arrays, we constructed elastic maps and their extensions in the space of first principal components. We analyzed changes in the cluster pattern of the studied volume of multidimensional data in different time periods. The changes occurring to individual selected clusters of words in different periods of time were considered.

Figure 2 shows an illustration of an elastic map extensions with data density rendering, plotted at 1-month intervals. You can see clusters of words - a large cluster formed in the middle of the left edge and two clusters on the right edge of the figure, located in the upper and lower corners. The figures reflect the changes that have occurred. The cluster located at the right edge of the figure in the upper corner is weakening. The large cluster located on the left edge is dropping closer to the middle. Thus, the elastic maps and their extensions reflect the changes in the multidimensional array of text data that occurred over a time interval equal to 1 month. The details of the changes occurring in the composition of individual clusters will be shown below.

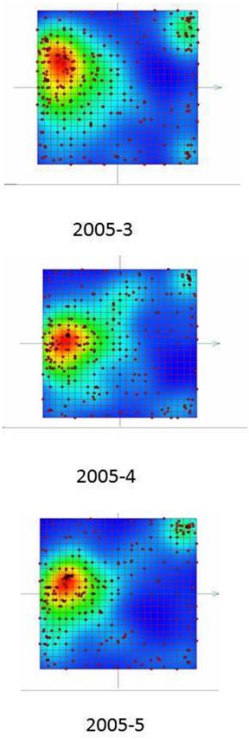

Figure 2. Evolution of the multivariate array under study on data obtained at an interval of 1 month.

The following Figure 3 similarly shows a picture for the evolution for the multivariate array considered at different points in time, with an interval of 1 year.

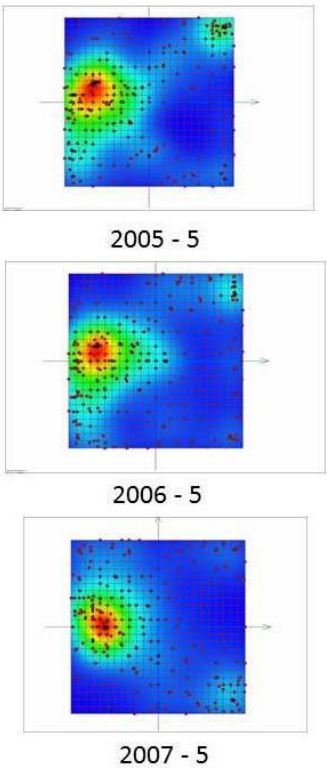

Figure 3. Evolution of the multivariate array under study on data obtained at an interval of 1 year.

Evolutions of the multidimensional array at different time intervals can be successfully shown using the visual representations shown in Figures 2 and 3. Such visual representations make it possible to trace changes in the overall cluster picture of the multidimensional array at different points in time.

3. EVOLUTION ANALYSIS FOR A SINGLE CLUSTER

In this section, we consider the changes occurring to individual clusters. To trace the evolution of individual clusters, we choose a characteristic cluster at the top of the right edge of the extension. As Figures 2 and 3 show, these clusters can be observed at all time points considered. Let us denote for the future the upper cluster of the right edge K1. Let us consider what happens to its qualitative composition at different points in time. For illustrative purposes, the cluster K1 will be highlighted with colored lines.

Figure 4 shows the view of the upper right edge for the multidimensional array under study at the time moment 2005-3. The yellow line in the figure highlights the resulting cluster K1. It consists of the following nouns: ВЛАСТЬ (POWER), РУКОВОДСТВО (LEADERS), СТОЛИЦА (CAPITAL), PARTY (ПАРТИЯ), ЛИДЕР (LEADER), ПРЕМЬЕР (PRIME MINISTER), ОППОЗИЦИЯ (OPPOSITION), ПАРЛАМЕНТ (PARLIAMENT), ПРАВИТЕЛЬСТВО (GOVERNMENT), СУД (COURT). Note that at this point in time 2005-3 an additional cluster is formed under cluster K1 with a group of words with semantic affinity: ВОЕННЫЙ (WARRIOR), ВОЕННОСЛУЖАЩИЙ (WAR-MAN), АРМИЯ (ARMY), СОЛДОТ (SOLDIER), РАЗВЕДКА (INTELLIGENCE SERVICE), СПЕЦСЛУЖБА (SPECIAL SERVICE), ГАЗЕТА (NEWSPAPER), УЧЕНЫЙ (SCIENTIST).

Figure 4. Clusters of the upper right edge for the multivariate array under study at time 2005-3.

Figure 5 shows a similar view of cluster K1 on the right edge of the extension at the time moment 2005-4. The composition of cluster K1 remains the same, but additional words are added to the words of the cluster: ПРОКУРАТУРА (PROSECUTOR'S OFFICE), БАЗА (BASE), ОРГАН (ORGAN,) ПОЛИЦИЯ (POLICE). At the same time, it should be noted that the cluster, located in the previous moment of time below the cluster K1, disappears. In its place remain the words: СПЕЦСЛУЖБА (SPECIAL SERVICE), СМИ (MEDIA), ГАЗЕТА (NEWSPAPER).

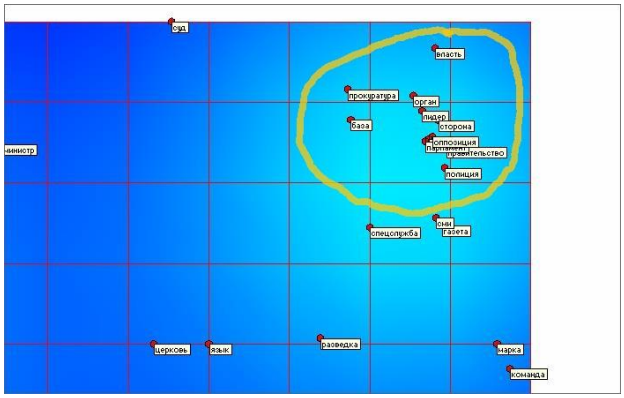

Figure 5. Clusters of the upper right edge for the multivariate array under study at time 2005-4.

The following figure 6 shows the results for the K1 cluster at time 2005-5. Cluster K1 at this point in time has the following composition: ВЛАСТЬ (AUTHORITY), СУД (COURT), СТОРОНА (SIDE), ПРАВИТЕЛЬСТВО (GOVERNMENT), ГОРОД (CITY), ВОЙСКО (TROOP), ПОЛИЦИЯ (POLICE), ПАРЛАМЕНТ (PARLIAMENT), ГАЗЕТА (NEWSPAPER), ПРЕЗИДЕНТ (PRESIDENT), ПРЕМЬЕР-МИНИСТР (PRIME MINISTER), ЛИДЕР (LEADER).
Note that just below cluster K1 in Figure 6 is a group of words: ТЕРРИТОРИЯ (TERRITORY), ГРАНИЦА (BORDER), РАЗВЕДКА (INTELLIGENCE), МИНИСТР (MINISTER), ЯЗЫК (LANGUAGE).

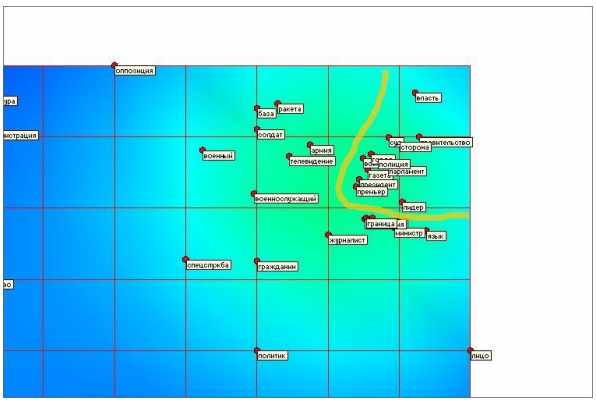

Figure 6. Clusters of the upper right edge for the multivariate array under study at time 2005-5.

Thus, applying the approach related to the analysis of snapshots of the elastic map allows us to successfully trace the evolution of cluster composition, where the interval between the snapshots is 1 month.

Now let us try to trace the evolution of clusters using snapshots taken at an interval of 1 year. We will consider the dataset obtained at the time of 2005-5, shown in Figure 6, as the initial dataset. Figure 7 shows a similar view of the clusters on the right edge of the sweep at time 2006-5. On the upper right edge of the extension of the elastic map two clusters are formed, which are highlighted in yellow in the figure. One of the clusters consists of the words: ВЛАСТЬ (POWER), ПРАВИТЕЛЬСТВО (GOVERNMENT), ГАЗЕТА (NEWSPAPER), КЛУБ (CLUB), ЯЗЫК (LANGUAGE), СТОРОНА (SIDE). The other cluster includes: ПАРЛАМЕНТ (PARLIAMENT), ПОЛИЦИЯ (POLICE), СТОЛИЦА (CAPITAL), ПРЕЗИДЕНТ (PRESIDENT), АРМИЯ (ARMY), СПЕЦИАЛИСТ (SPECIALIST), ВОЕННЫЙ (MILITARY), БАЗА (BASE), ВОЙСКО (ARMY). It seems that at this point in time cluster K1 of the original array divides into two clusters, since both include words from the original cluster.

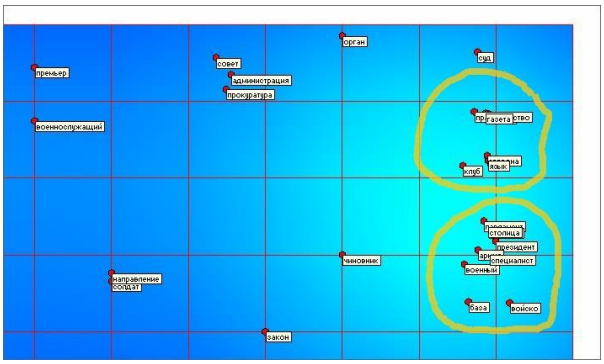

Figure 7. Clusters of the upper right edge for the multivariate array under study at time 2006-5**.**

Figure 8 shows the results for time point 2007-5. At this point in time, the cluster of the upper right edge weakens even more. At this point in time, it includes only 4 words: РЫНОК (MARKET), БАНК (BANK), ОРГАНИЗАЦИЯ (ORGANIZATION), КОМПАНИЯ (COMPANY). At a distance from the cluster, one can distinguish the words СТРАНА (COUNTRY), ГОСУДАРСТВО (STATE). Also at a distance you can distinguish the word ПАРТИЯ (PARTY). Recall that at previous times these words were included in the K1 cluster.

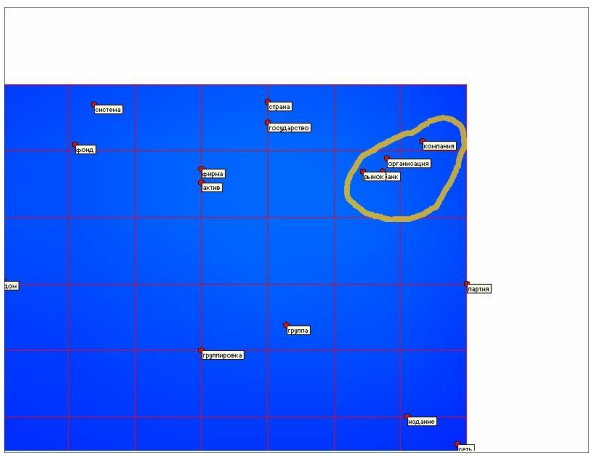

Figure 8. Clusters of the upper right edge for the multivariate array under study at time 2007-5**.**

Thus, the results of numerical experiments on real multidimensional arrays of textual information have shown the possibility of analyzing the dynamics of changes in the composition of individual groups of semantic proximity (clusters) and their quantitative characteristics according to data obtained at different time intervals.

As a result, the method of computational and visual analysis of multidimensional dynamic arrays of textual information can be formulated as a chain of sequential operations:

1. Construction of the initial multidimensional data at different points in time to obtain the studied dynamically changing amount of data.

2. Construction of elastic maps for each moment of time through a certain time interval.
3. Construction of developments of elastic maps on the plane formed by the first two principal components at each moment of time.
4. Obtaining a picture of the evolution of the general cluster structure of the studied dynamically changing amount of data over time.
5. Selection of individual clusters of interest to the user.
6. Selection of each individual cluster for each moment of time on the development of an elastic map. Tracing the evolution of each individual cluster of interest.
7. Description of the evolution for each individual cluster of interest.

4. CONCLUSIONS

To analyze the "visual portrait" of a multidimensional data volume and to study the cluster structure in multidimensional volumes, the technology of constructing elastic maps is currently effectively used. Technologies for constructing elastic maps are methods for mapping points of the original multidimensional space onto manifolds of lower dimension embedded in this space. A number of author's studies were devoted to the visual analysis of the cluster pattern in multidimensional text volumes.

However, all previous studies have dealt only with static data volumes, that is, volumes that do not change over time. At the same time, the most active research attention is now paid to the visual analysis of dynamic data volumes - data that changes over time. The study of changes occurring in the studied multidimensional volumes of textual information will help to identify the factors that cause such changes. Such a study should be subject to both the visual cluster picture of the studied volume as a whole and individual clusters.

Previously, the authors presented a hypothesis that elastic maps and their extensions can be an effective means of visual analysis of dynamic arrays of multidimensional text data by treating elastic maps extensions as snapshots of the cluster structure of a dynamic data array presented at different points in time. Approbation of this approach was previously carried out on a set of artificial data that simulates changes in the original data set at different points in time.

The purpose of this work was to analyze data on real volumes of textual information received at different points in time. The data was considered as snapshots obtained at time intervals of different lengths. For all moments of time, elastic maps and their extensions in the space of the first principal components were calculated. The analysis of changes in the cluster pattern of the studied volume of multidimensional data has been carried out. Changes occurring with individual clusters are considered.
Based on the results of numerical experiments, it can be argued that the proposed visual analysis approach is workable and can serve as a working tool for studying time-varying arrays.

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., 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., 2021. Visual analysis of text data collections by frequencies of joint use of words, ISPRS Archives, XLIV-2/W1-2021, 21–26, 2021.

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.

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 2023).

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]