Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 329–335

DOI: 10.15439/2015F50
ACSIS, Vol. 5

Mapping Evaluation for Semantic Browsing

Veslava Osinska
Institute of Information Science and Book Studies, Nicolaus Copernicus University, ul. Bojarskiego 1, 87-100 Toruń, Poland
Email: wieo@umk.pl

Adam Jozwik
Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Science, Warsaw; Institute of Computer Science, College of Social and Media Culture, Toruń, Poland
Email: adamj346@wp.pl

Grzegorz Osinski
Institute of Computer Science, College of Social and Media Culture, ul. Starotoruńska 3 87-100 Toruń, Poland
Email: gos@fizyka.umk.pl

***Abstract*—The paper contributes to the problem solving in semantic browsing and analysis of scientific articles. With reference to presented visual interface, four – the most popular methods of mapping including own approach - MDS with spherical topology, have been compared. For a comparison quantitative measures were applied which allowed to select the most appropriate mapping way with an accurate reflection of the dynamics of data. For the quantitative analysis the authors used machine learning and pattern recognition algorithms and described: clusterization degree, fractal dimension and lacunarity. Local density differences, clusterization, homogeneity, and gappiness were measured to show the most acceptable layout for an analysis, perception and exploration processes. Visual interface for analysis how computer science evolved through the two last decades is presented on website. Results of both quantitative and qualitative analysis have revealed good convergence.**

I. Introduction

Nonlinear growth of scientific writing imposes a new forms of academic databases management. The latter includes the both retrieval and analytical exploration. Analysts, science of science professionals, science policy makers need various computing, statistical and visualization tools to monitor how science or scientific domains evolved.

Authors designed and described in a series of papers [1], [2] the visual interface for analysis of dynamics of computer science through the two last decades[1]. Screenshot of spherical application is shown on Figure 1. Users can interact, manipulate and browse the data and see how graphical pattern change in time. The nodes represent scientific articles from digital library and the colour – appropriate thematic category. Similarity metrics was based on semantic relations between documents [2]. In order to generate 3D layout, multidimensional scaling (MDS) technique was applied and enriched by Morse potential [1].

Overlapping spots show where the categories mutually integrate that means an articles at that location are semantically similar. Visualization of classified documents reveals both organization of digital library content as well as allows users to track how it changes over time. This paper presents further study on visualization maps. Authors decided to test this prototype regarding to mapping algorithms. Four mapping methods were compared in terms of dynamics and analytical possibility of output visualization. The next chapter shows the outline of VxOrd, MDS, VOS, SOM as the most popular methods.

II. Mapping preview

How we, as analysts, perceive and understand the connections between data, depends on graphical layout. Thus, the final structure of visualized knowledge can be drawn either by spatial arrangement (2D or 3D) of analysis units or by the relationship between nodes in graph or combination of these two.

One of the basic ordination algorithm - VxOrd extends a traditional force-directed approach [3]. VxOrd determines the both number and size of clusters automatically based on the data. Popular software for data mapping and visualization - Gephi[2] uses this technique. Due to Gephi users can analyse large networks consisting of even millions of nodes. The most popular technique for dimension reduction is MDS, which involves minimizing the difference between Euclidean and graph-theoretic distances. MDS has been widely applied for constructing knowledge maps of authors, articles, journals, and keywords [3]-[5]. The same satisfactory representation of knowledge can be produced by use of new mapping technique VOS introduced in series of works by Van Eck and Waltman [6],[7]. The idea of VOS is to minimize a weighted sum of the squared distances between all pairs of items.

Dimension reduction can be also achieved in self organized maps – the kind of unsupervised neural network which aimed to project high-dimensional data into a lower-dimensional space [4]. The nodes (input vectors) form two dimensional regular grid; node's neighbourhood is defined to be all connected nodes. During training process similar input vectors stimulate adjacent neurons and therefore output SOM map shows semantic relationships between data, where similar items are mapped close together. Comparing MDS and VOS,

[1]http://www-users.mat.umk. pl/~garfi/vis2009v3/

[2]http://gephi.github.io/

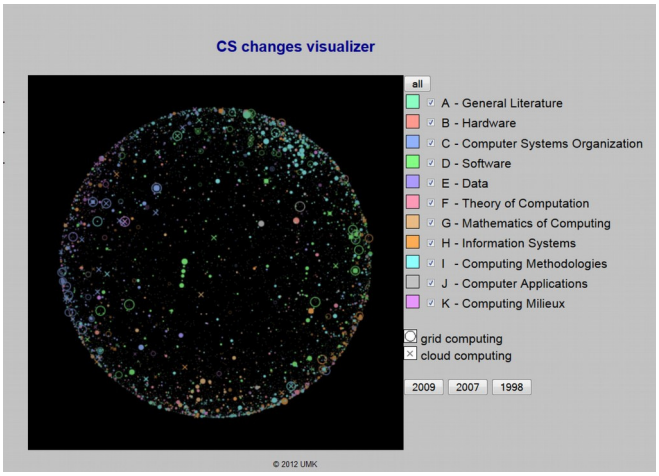


Fig. 1 Interface screenshot - the prototype of application for semantic browsing

researchers concluded that maps constructed using VOS approach provide a more satisfactory representation of the underlying dataset [8]. MDS-VOS tests revealed the first one is distance preserving while the second - topology preserving technique [9],[10].

III. Methodology

The complex structure of large graphs commonly is measured by modularity. According to Fortunato [11], modularity can be defined as the function which evaluates the goodness of partitions of graph and is defined by ties between vertices, vertices and hubs. Structure described by linked objects does not match with character of present data. Authors analyse the visualization of classified documents. Primary categories are taken from library classification. All relationships between collected classes and documents were predefined by specialists permanently working on computer science taxonomy. Visualized articles are assumed to reveal semantics while keeping their thematic similarity. These new correlations (down-top) between data allow to rich independence of the original organization of items (top-down). As every rigid scheme, it is characterized by adequacy and disjointedness of subclasses. Therefore, evaluation metrics can be based on spatial configuration of visual layout instead of links outline. To analyze the final nodes distribution, image processing methods were applied. Then evaluation of graphical pattern should be carried out taking into account the accuracy, topology (space filling and capacity) and perception abilities of users.

The authors present alternative approach: a sphere surface has been selected as a target mapping space. There are some arguments for a sphere surface: it “has no edges and therefore it is possible to represent not only local similarities but also large-scale ones regarding the whole space. The benefit of a curved surface in comparison to a plane one is a more capacious exploration space” [1,10]. 3D visualization is a popular but also challenging method in large dataset mapping and modelling.

A. Assumptions

By implication, evaluation process will touch how to fit interface to the requirements of analysts and domain experts. The study is based on the following assumptions.

1. A given visualization layout might serve as a graphical interface for the exploration and semantic retrieval of scientific articles. From this point of view the most important is configuration on the bottom level (documents) – then can be evaluated the spatial distribution of nodes.
2. Current modifications by editors of the original classification are aiming at its improvement. The classification reflects the most of current changes in computer science. Quickly developing categories will form dense clusters and overlap each other. These tendencies must be visible on visualization maps generated for different time periods.
3. In the construction of the ergonomic user interface, such features as capacity, homogeneous distribution and edgelessness must be taken into consideration.

Short movie[3] shows how three dimensional configuration allows the user to analyze semantic distribution of articles and its behavior in time.

B. Evaluation steps

On Figure 2 we can see the elements of evaluation process. Continuity characteristics is crucial for present study. For this purposes clusterization potential was validated by machine learning and image recognition algorithms. Structural complexity can be evaluated by fractals analysis. Quantitative measures are different for several maps and the changes tendency are essential too.

All dimension reduction methods determine the arrangement of classes and subclasses nodes. Documents distribution was calculated by using geometrical rules in 2D or 3D space [10]. Obtained pattern became the basic material for comparison and further study.

If we plan to involve users to scientific domain analysis, visualization interface must be user-friendly and carry good navigational features. Another usefulness of such application is retrieval of semantically similar documents. Precision in this case will be an appropriate measure of this visual searching system.

C. Research material

Visualization maps were obtained by using the same data but distinct in terms of data configuration (like matrixes versus data pairs), mapping algorithms and space topology. The series of every ten-year layouts show the changes of pattern and thus the evolution of the ACM classification and computer science knowledge (see Appendix). An insight into the differences in graphical patterns could reveal the most and the least complex structures due to human perception. The system of human perception is able to recognize a natural tex

[3]www.wizulizacjainformacji.pl/unas/interface.avi

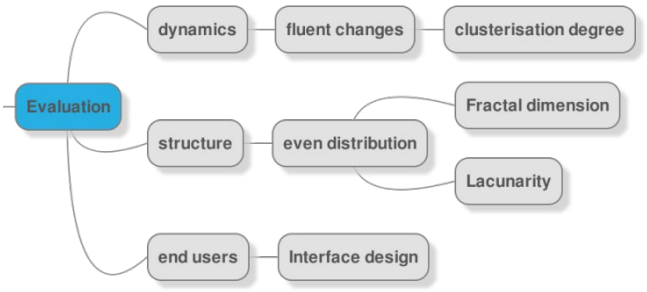


Fig. 2 Decision map of visualization evaluation steps

ture appearing in nature as a result of evolutionary adaptation. The human vision allows determining approximately whether the perceived structure differs one from another in terms of complexity[12].

For example, the first series of maps is characterized by a relative even distribution while VxOrd by data grouping on edges and bends. Furthermore, VxOrd and VOS distributions are highly limited to the output geometry [10]. The result may be a non-effective space for navigation in those cases. But human perception cannot be one of the main criteria for comparison and estimation of visual layouts, although useful in the final conclusions. Quantitative approach requires that the authors analyze local density differences and quantify clusterization, rarefaction, homogeneity and porosity.

Output maps can be described by both density and colour of nodes. If information about the main thematic category assignment is excluded (i.e. the colour), the clusters can be identified by density only. Consequently, clusterization and its changes can deliver information on how knowledge advances and how knowledge organization changes throughout two decades, independently of the primary (original) classification.

IV. Quantitative Analysis of Maps

A. Clusterization and its dynamics

To identify clusters on given maps we used the most popular partition technique, based on distances between points and/or points to centroids – k-means clustering [11]. Algorithm aims to minimize the within-cluster sum of squares:

$$\sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - c_i\|^2 \qquad (1)$$

where S_i indicates the subset of points in the i-cluster, c_i – the centroid of cluster.

The disadvantage of this algorithm in our assignment is the requirement to know the number of clusters. To find optimal number of classes we modeled the data by a set of Gaussian distributions [13].

The number of components can be estimated by the Bayesian Information Criterion [14] (BIC), which is based on a penalization of the observed log-likelihood – the function of x, θ. The preffered model is the one with the lowest value of BIC which decides about the number of clusters. Thus the 6 clusters are recognized for the MDS –sphere (authors method) and SOM and 8 for both VxOrd and VOS. These new clusters reorqanize initial data assignment to the 11 main initial categories, coded by colour. By k-means clustering the centroids of clusters are found, demonstrated on Figure 3 according to the the MDS –sphere map.

Dynamical characteristics of clustering are crucial for a final evaluation of the presented approaches in terms of structural analysis. In any sequence of maps it is possible to find the one with a highly developed clusterization just intuitively.

To evaluate clustering and its dynamics, a misclassification rate offered by the standard k nearest neighbour (k-NN) rule was used as a criterion. That error rate was estimated by the leave-one-out method [15],[16]. The k-NN rule assigns the classified object to the class most heavily represented from among its nearest objects in the training set (i.e. nearest neighbours). The reference set, also called a training set, is a set of objects with a known class membership and in a certain sense it defines the considered classes. The leave-one-out method consists in the classification of each object from the reference set by the decision rule obtained from the training set decreased by the currently classified object. The ratio of the number of misclassified objects to a numerical force of the reference set estimates the above mentioned error rate used as the clustering quality criterion.

The low error rate value denotes that the considered classes (or clusters) differentiate easily, but high values of the error rate mean that the classes overlap. The leave one out method is very convenient in the case of classifiers based on the k-NN rule since no training is required. This property of the k-NN classifiers was intesively used for creating a fuzzy k-NN rule proposed by one of the authors of the present work [17] and for introducing the more sophisticated pair-wise k-NN classifier [18].

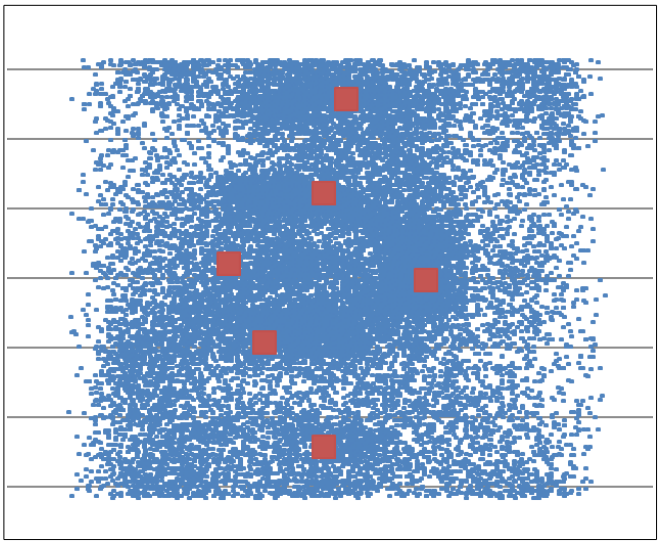


Fig. 3 Data distribution with six clusters centroids (along "horseshoe" shape).

TABLE I.EVALUATION OF CLUSTERIZATION BY K-NEAREST NEIGBOURS METHOD FOR EACH VISUALIZATION MAP BASED ON ERROR RATES.

Phase	Training			Testing		
	1988	1998	2009	1988	1998	2009
Authors' method	0.0140	0.0115	**0.0074**	0.8890	0.8300	-
VOS	0.0119	**0.0078**	0.0110	0.9011	-	0.8710
VxOrd	0.0084	0.0107	**0.0045**	0.8707	0.7898	-
SOM	0.0143	**0.0121**	0.0128	0.8480		0.8090

In every series of visualizations it is possible to point at the map with the clearest clustering structure – model map. The other, as it can be assumed, develops towards clustering pattern. If the algorithm for pattern (points) clustering of model map is trained, other ones can be tested by using the nearest neighbour method. Which one serves as a training set and which one as a testing set can be found first by evaluation of the standard k-NN classifier. The lower error rate means a better clustering structure (bold numbers in *Training* part of Table 1).

The results of training allow selecting an apropriate dataset (bold) for testing. During the testing phase (*Testing* part) the clusterization quality can be tracked on the basis of error rate changes. It is worth noting that comparison must be made along rows, not columns, because of the use of different methods to generate patterns.

The data in Table 1 show that the clusterization increases in one case – the first row, which characterizes the authors' approach. It proves continuous changes of CCS towards overlapping categories and reorganization needs.

B. Even distribution, FD and Lacunarity

The authors' study of visualization maps relates to visual interface: which method of mapping can deliver the best way to explore a complex dataset of scientific articles?

Restrained homogenous distribution on a sphere surface can be estimated by the volume of empty places. The more holes in the pattern, the more heterogeneous it is. The appropriate parameter is *Lacunarity* - the degree of holes distribution having the lowest value for indeterminate structure. Lacunarity is often used in medical imaging for detection of structural changes in bone texture on radiographs [19].

Visualization maps can be considered as flat textures associated with the patterns of documents nodes distribution. Lacunarity λ is defined as:

$$\lambda_{\epsilon,g} = CV^2_{\epsilon,g} = \left(\frac{\sigma}{\mu}\right)^2_{\epsilon,g} \quad (2)$$

where σ is the standard deviation and μ is the mean for pixel per box at this size ε, in a box at this orientation g. Lacunarity pertains to both gaps and heterogeneity. To simplify, the more *gappiness* in the image (i.e. sparsely occupied maps), the higher lacunarity. Some recent research has shown that there is a correlation between lacunarity and *fractal dimension*, FD [20],[21].

The FD is a complexity indicator with a non-integer value. The fractal dimension could be characterized as a scale of transition to homogeneity and is therefore very practical in the dynamics study case. Because the maps were generated by three different mapping algorithms they present distinct homogeneity, what is according to our assumption, one of the criteria of the ergonomic visual interface.

The values of lacunarity and the FD for every map are shown in Table 2. The highest value in each row (bold numbers in the first part of Table 2) indicates a map with the large porosity (gappiness). Bold FD values in the second part of the same table means the best formed structure (implicitly clear clusterization). The first row data presents a continuous growth of complexity degree with simultaneously dense occupation (high FD and lowest lacunarity values). Other (VOS, SOM and VxOrd) demonstrate oscillations are difficult to interpret. Consequently, the dynamics of each index across time for every method was evaluated. Lacunarities of every method should not be compared because of different spanning geometry.

If should be taken into consideration that the fractal dimension for random (or pseudo-random) distribution equals 2.77, the more does the FD tend to this value, the pattern is more homogeneous [22]. And inversely, the low FD (bold numbers in Table 2) means the distribution resembles linear. A stable structural change (in contrary of step change) in time is proved by the authors' method.

V. DISCUSSION

According to authors' conception, in order to measure dynamics of graphical patterns we need to focus on how complexity evolves. Therefore clustering resolution has been tested by use of machine learning and pattern recognition algorithms.

TABLE II. LACUNARITY AND FD FOR EACH VISUALIZATION MAP.

Method	Lacunarity			Fractal Dimension		
	1988	1998	2009	1988	1998	2009
Authors' method	0.0185	0.0155	**0.0147**	2.34	2.39	**2.50**
VOS	**0.0035**	0.0144	0.0067	2.23	2.15	**2.40**
VxOrd	**0.0247**	0.054	0.0421	2.18	2.15	**2.23**
SOM	0.3340	**0.305**	0.319	1.82	1.84	**1.97**

The authors proposed the qualitative measures for evaluation of structural changes of pattern: FD and lacunarity.

In general, we can conclude: the larger complexity degree, the lower randomness. On the other hand, the complex structure is also can be determined by the clustering level. The current findings confirmed by presented measures (Table 1, Table 2) show that clusters have become more explicit with time at the maps generated by authors' approach and at the same time tend to uniform distribution (lowest lacunarity and FD resembles the value of random distribution). The high lacunarity informs about dense network of holes among others, due to overlapping pattern. Next acceptable technique in the terms of changes continuity is SOM. Qualitative approach to compare visualization maps [10] shows the similar results: both MDS-sphere and SOM reveal consequence in dynamics changes, moreover VOS and VxOrd – inappropriate topology for data exploration [10, Appendix]

How easy users can play with data and analyze their change – show the ergonomic properties of visualization interface. Homogeneous occupation of visual layout, edgeless, continuity in changes should feature good visualization [23]. Several papers described particularly this application from the end-users-analysts point of view [10], [24]. Another practical aspect pertain relevant documents retrieval due to visual representation. This still remains the main direction of current study. After receiving good precision for a small sample, authors intend to repeat experiment with bigger dataset and all presented methods.

Recent research [20],[21] show a strong correlation between the FD and lacunarity. To check this, it is required to have a more representative dataset i.e. be multi-various. To find essential changes including paradigms, the period of analysis must be extended to three or four decades, i.e. until 2017. There basic technological problem has appeared: the ACM has changed the classification and applied it to the collection of 2013. The success to supplement the dataset depends on whether the ACM will standardize the old classification schemes according the new version and adapt it to the whole dataset.

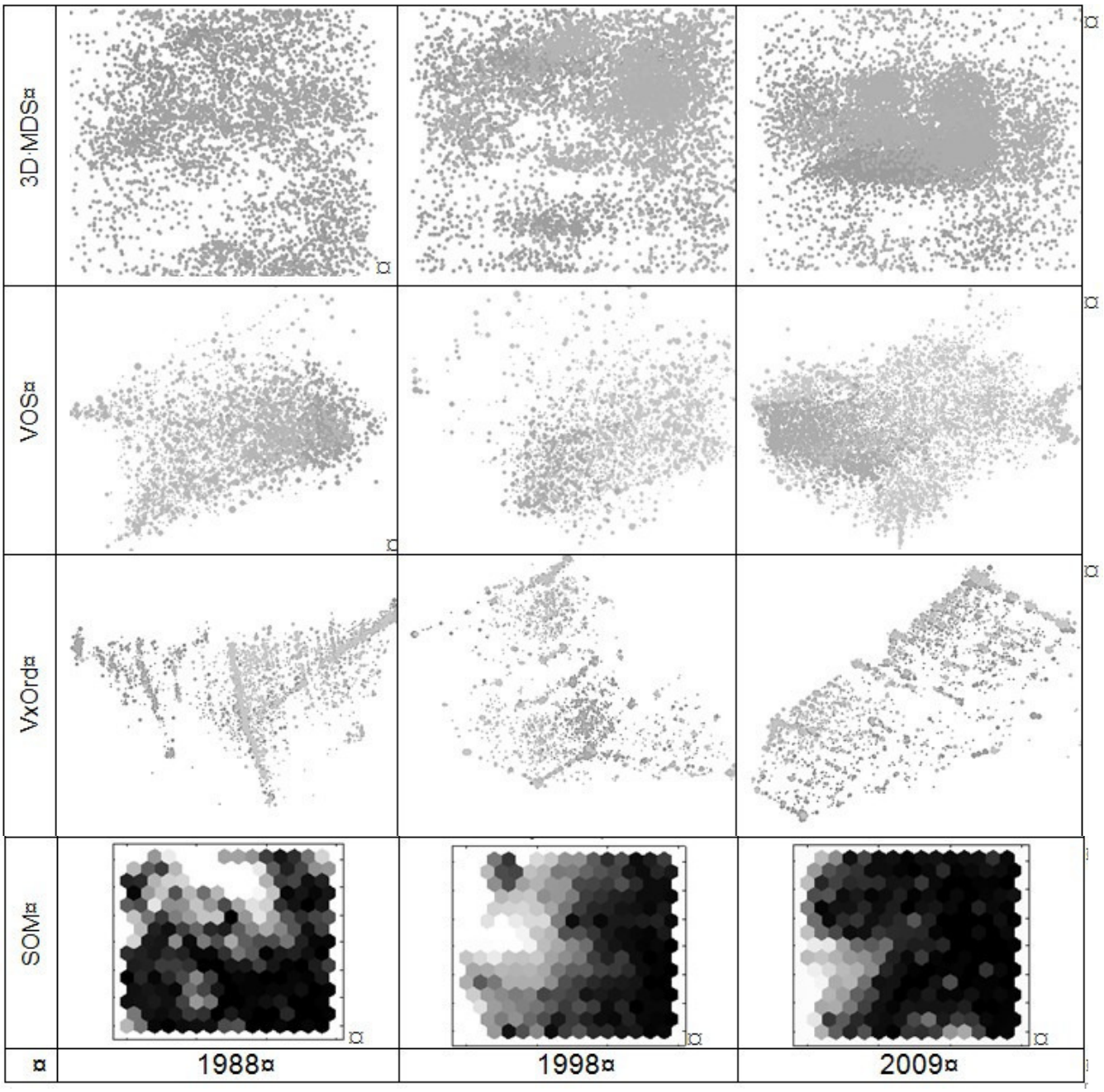
Undoubtedly, a sequential series of three maps is not enough to estimate knowledge evolution dynamics. A more multi-variety dataset to track all changes in fast growing knowledge is needed, but truly objective circumstances concerning data gathering were appeared. However, proposed measures can be considered if we need to select the best data distribution in the terms of interface functionality.

VI. SUMMARY

Visual interface for analyzing how computer science evolved through the two last decades is briefly presented in current paper. This application includes an interactive 3D map of scientific articles organized by their semantic relationships. The authors proposed the conception how to quantitatively evaluate different visualization maps in respect of possibilities of dynamics analysis. They also characterize topological arrangement in the terms of navigation functionality.

Four methods of mapping including own approach of mapping (MDS with spherical topology) have been compared. Quantitative measures allowed selecting the most appropriate mapping way with an accurate reflection of the current changes of computer science. In the quantitative analysis authors tracked the changes of pattern clusterization over time. Clusterization degree they evaluate using machine learning and pattern recognition algorithms (Table 1). They adopted both lacunarity and the fractal dimension of visualization patterns to find the scale of randomness in dynamics (Table 2). Moreover the local density differences, clusterization, rarefaction, homogeneity, and gappiness were measured to show the most acceptable layout for analysis, perception and exploration processes. 3D MDS maps (authors' approach) and SOM have shown the better properties than VOS and VxOrd. These results have proved the findings and interpretations obtained from qualitative analysis [9]. Given maps have revealed essential changes in computer science literature during the time of the development of the CCS classification compared.

APPENDIX



References

[1] V. Osinska and P. Bala, "Classification Visualization across Mapping on a Sphere", in: *New trends of multimedia and Network Information Systems.* Amsterdam: IOS Press, pp. 95-107, 2008. ISBN 978-1-58603-904-2.

[2] V. Osinska, P. Bala and M. Gawarkiewicz, "Information Retrieval across Information Visualization". IEEE Xplore Digital Library: *Proceeedings of 2012 Federated Conference on Computer Science and Information (FedCSIS),* Wroclaw, 2012, pp. 233 – 239.

[3] K. W. Boyack, B. N. Wylie and G.S.Davidson. "Domain visualization using VxInsight for science and technology management". *Journal of the American Society for Information Science and Technology,* 53(9): 764-774, 2002. doi: 10.1002/asi.10066.

[4] Ch. Chen, *Information Visualization. Beyond the Horizon.* 2nd ed. London: Springer, 2006, pp.143-170. ISBN: 978-1-84628-579-0.

[5] K. W. Boyack, R. Klavans and K. Börner, "Mapping the backbone of science", *Scientometrics,* vol. 64(3): 351-374, 2005. doi: 10.1007/s11192-005-0255-6.

[6] N. J. Van Eck and L. Waltman, "VOS: a new method for visualizing similarities between objects", in *Advances in Data Analysis: Proceedings of the 30th Annual Conference of the German Classification Society* (eds HJ Lenz, R Decker), London: Springer, pp. 299-306, 2007

[7] N. J. Van Eck and L. Waltman, "How to normalize cooccurrence data? An analysis of some well-known similarity measures", *Journal of the American Society for Information Science and Technology,* 60(8): 1635-1651,2009. doi: 10.1002/asi.21075.

[8] N. J. Van Eck, L. Waltman, R. Dekker and J. Van den Berg, "A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS". *Journal of the American Society for Information Science and Technology,* 61(12): 2405-2416, 2010.

[9] F. Moya-Anegón, V. Herrero-Solana and E. Jiménez-Contreras. "A connectionist and multivariate approach to science maps: the SOM, clustering and MDS applied to library and information science research". *Journal of Information Science,* 32(1): 63-77, 2006. doi:10.1177/0165551506059226.

[10] V. Osinska and P. Bala, "Study of dynamics of structured knowledge: Qualitative analysis of different mapping approaches", *Journal of Information Science,* 1-12, 2014. doi: 10.1177/0165551514559897.

[11] S. Fortunato, "Community detection in Graphs", *Physics Reports,* 486: 75-174, 2010. doi: 10.1016/j.physrep.2009.11.002.

[12] C. Ware, *Information Visualization: Perception for Design.* CA: Morgan Kaufmann, pp. 11, 188, 273, 2004. ISBN 0123814642.

[13] J. D. Banfield and A.E. Raftery AE, "Model-based gaussian and non-gaussian clustering", *Biometrics,* 49: 803-821, 1993. doi: 10.1093/biomet/63.3.413.

[14] C. Biernacki, G. Celeux and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22: 719-725, 2000.

[15] P. A. Devijver and J. Kittler, *Pattern recognition. A statistical approach,* London: Prentice Hall, 1982.

[16] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern classification,* New York: John Wiley & Sons, 2001.

[17] A. Jozwik, "A learning scheme for a fuzzy k-NN rule", *Pattern Recognition Letters,* 1: 287-289, 1983. doi: 10.1016/0167-8655(83)90064-8.

[18] A. Jozwik, S. Serpico and F. Roli, "A parallel network of modified 1-NN and k-NN classifiers -application to remote-sensing image classification", *Pattern Recognition Letters,* 19: 57-62, 1998.

[19] R. E. Plotnick, R. H. Gardner and R. W. O'Neill "Lacunarity indices as measures of landscape texture", *Landscape Ecology,* 8(3): 201-211, 1993.

[20] A. Forsythe et al., "Predicting Beaty: Fractal dimension and Visual complexity in art", *British Journal of Psychology,* 102, 49-70,2011. T. G. Smith, G. D. Lange and W.B.Marks, "Fractal Methods and Results in Cellular Morphology", *Journal of Neuroscience Methods,* 69: 1123-126, 1996. doi: 10.1016/S0165-0270(96)00080-5.

[21] V. Osinska, "Fractal analysis of Knowledge Organization in Digital Library", in Katsirikou A, Skiadas CH (eds) New Trends in *Qualitative and Quantitative Methods in Libraries,* Singapore: World Scientific Publishing, pp. 17-23, 2011.

[22] W. A. Pike et al., "The Science of Interaction", *Information Visualization,* vol. 8, 4: pp. 263-274, 2009.

[23] V. Osinska, J. Dreszer-Drogorob, G. Osinski and M. Gawarkiewicz "Cognitive Approach in Classification Visualization. End-Users Study", in *Classification & Visualization: interfaces to knowledge* (ed A. Slavic et al), Hague, Holland, 23 -25 October 2013, Würzburg: Ergon Verlag, pp. 273-283. ISBN 978-3-95650-007-7.