Fractal Analysis of Knowledge Organization in Digital Library

Veslava Osinska[1]

Institute of Information science and Book Studies, Nicoulaus Copernicus University
Toruń, Poland, e-mail: wieo@umk.pl

Abstract: Visualization of the large-scale collections of information became one of the essential purpose in data analysis. The new methods of visualization are increasingly applied as a significant component in scientific research. Particularly qualitative nature of Infoviz studies (Information visualization) can be combined with quantitative character of digital libraries volumes. This paper describes and demonstrates the case of hierarchical structure visualization i.e. visual representation of both classification adopted by ACM (Association for Computing Machinery) digital library and classification universe. Given maps were processed by nonlinear graphical filters. Finally fractal dimension (FD) and derived techniques have used to analyze the patterns of clusters on the visualization maps. Quantification of output graphical representation by means of fractals makes possible to adjust visualization parameters as well as evaluate initial classification scheme and its dynamical characteristics.

Keywords: fractal analysis, fractal dimension, visualization, classification scheme, knowledge mapping ,

1. Introduction to fractal analysis

In analysis of large datasets of digital libraries advanced numerical methods became well-established. It is possible to draw two main approaches in the processing information in library collections. The fist one, so-called conventional is measuring of quantitative characteristics of library database such as number of records, bibliographic data and its dynamical changes. Knowledge of statistical methods in this case is fundamental. Discovered correlations are usually presented in linear way by the tables, diagrams and charts. Another group of techniques lead to find nonlinear dependences between the objects. Non-linearity takes place when we try to describe the unstructured and inhomogeneous data, for example user-based Web 2.0 data. Mapping as common technique in information visualization (Infoviz) provide such complex dataset with nonlinear representation. Infoviz methods can reveal hidden structure of scientific data derived from bibliographic databases (Chen 2006, Börner 2003). By this way constructed visualization maps contribute to a better understanding of the knowledge organization and their dynamics as well as monitoring the scientific output overall. By mapping subject classification in Computer Science domain on a sphere surface it is possible to analyze the development of this dynamic field (Osinska&Bala 2008, 2010). Given visualization maps were processed by selected image processing methods, that is presented in current paper. Inhomogeneous distribution of documents nodes showed some latent structure in a pattern. Describing the data within such complex patterns could be solved by means of fractal analysis. Fractal is was coined by Mandelbrot and defined as "a rough or fragmented geometric shape that can be split into parts, each of which is (at least approximately) a reduced-size copy of the whole" (1982). This is the main feature of fractals and called self-similarity. When we magnify the patterns that are Euclidean, we look more and more details which recur in each level. Objects in nature can be approximated by fractals, for example: clouds, mountain ranges, frost crystals, snow flakes, fern leaves, various vegetables (cauliflower and broccoli).

Practically second significant feature of fractal objects such as fractal dimension is used in fractal analysis. This distinguishes fractals from Euclidean objects, which have integer dimensions. As a simple example, if we magnify a length of a square's side two times its area will increase four-times. The same operation in fractal case causes area changes less than 4 times. Fractal dimension (FD) is non-integer value, usually a smaller than topological dimension of proper primitive figure, thus it determines how fractal differs from Euclidean objects. Fractal dimension measures the degree of fractal boundary fragmentation or irregularity over multiple scales. carries important information about how a fractal fills a space where it is embedded. Another fractal's measure used in current work – lacunarity shows how a fractal fills space and is applied to further classify fractals and textures which, while sharing the same fractal dimension, appear very visually different (Mandelbrot 1982, Plotnick&Gardner 1993).

Fractal dimension of regular figures are the same as topological. For example 1,2,3 for line, square and cube respectively. Some instances of fractal dimension are quoted below[1]. FD for Koch snowflake equals 1.26, cloud – 2.5 , Norway coastline – 1.52, cauliflower - 2.66, human brain – 2.79, Tree - 2.7.

For fractal processing of visualization maps fractal analysis toolbox FracLac was used. FracLac is free software with user-friendly and intuitive interface; one can use it directly to perform many tasks in signal and images processing, including estimation, detection, modelling, classification, and so forth.

2. Classification mapping on a sphere

Research work consists of the visualization and analysis of documents classified by ACM Computing System Classification. Collection of abstracts is accessible in ACM Digital Library so own application allowed to gather metadata of articles. Classification tree contains three levels. The upper one consists of 11

[1] List of fractals by Hausdorff dimension. In: *Wikipedia. The Free Encyclopedia* [on-line]. Available online at URL: http://en.wikipedia.org/wiki/List_of_fractals_by_Hausdorff_dimension

main classes coded by 11 capital letters (from A to K). For the more precise categorization every article besides main classification is ascribed (in general by authors and/or editors) to one or more additional. Thus such common for different classes and subclasses documents can be considered as a measure of their thematic similarity. The innovative idea relies on estimation of co-occurrences of classes i.e. counting of common documents for every pair classes and subclasses, that result in construction of classes similarity matrix.

Osinska and Bala (2008, 2010) describe in detail the construction of a new graphical representation of original classification scheme in the 3D space, namely sphere surface. The final number of all possible classes and subclasses in collection was 353. Among all (sub)classes nodes, documents positions on a sphere were calculated from topological relations between main and additional classifications. Three variants of weights: 06:0.4, 0.7:0.3 and 0.5:0.5 were tested. Apparently fractal characteristics are helpful in qualitative comparison of obtained maps as well as selection the proper configuration. Figure 1 represents classes visualization on a sphere using 3 attributes: colour to indicate main class, intensity – tree level and a size – population of (sub)class. The documents nodes were coloured by their main class color. For convenient analysis cartographic projections of visualization layouts were used.

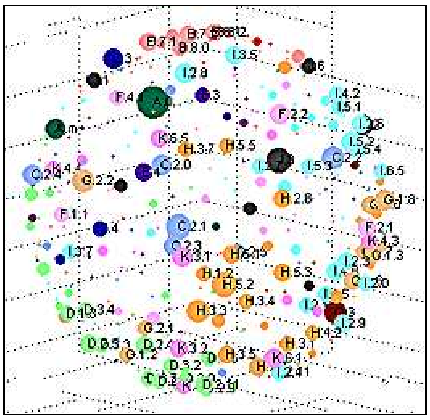

Figure 1. Classes visualization on a sphere.

3. Image set for analysis

By observing natural objects it is easy to notice most fractals have dimensions in the range [1,2], thus these dimensions are bigger than in flat figures case, but less than in solids. Therefore such fractals are some formation between straight lines and flat figures. Natural objects what we see around us (mountains, trees, clouds) are rather three-dimensional and have FD value converged to the value above 2. The object's topology is more complicated, the fractal dimension is closer to number 3.

If a low-resolution image is a small size it perceived by users as one with a good quality and sharp edges because of visual perception feature to focus vision within a limited field of view (Ware 2006). In the case of big size pictures the process undergoes blurring. Webmasters well know this effect and put into web galleries good quality miniatures of images served as links to the larger originals.

Visualization maps as a result of sphere surface projections on a plane were prepared according fractal analysis requirements. They need to be converted to the shades of gray and scaled to one universal size exact to a one pixel. To optimize the computation time three graphic formats TIFF, JPG and BMP were tested. Finally files exported as TIFF type and used for further research.

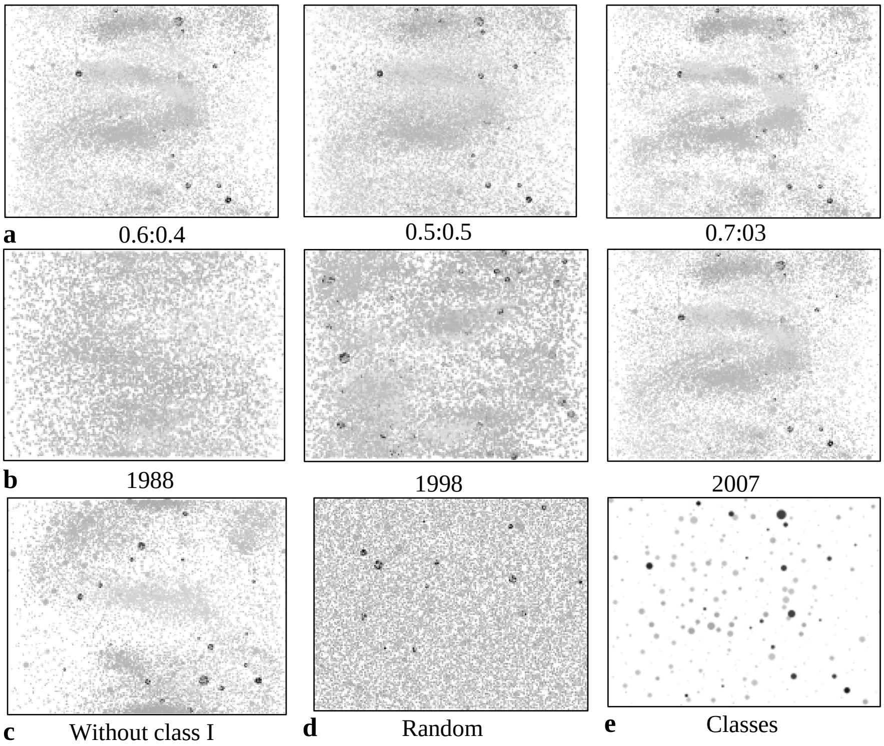

Figure 2. Comparison of given visualization maps after desaturation: a) for different proportions of the primary and additional classes in documents classification; b) for different publishing years; c) for modified classification after removing the class I; d) random distribution's map; e) map of all classes.

It is better to compare fractal structures visually if proper illustrations after desaturation[2] are all set into one observation view. This situation is presented on a Figure 2: classification maps for different proportions of the primary and additional classes (a), maps of articles published in different years (b), map for artificially changed original classification (c), control map of random distribution of the same quantity of points (d) and a map of 253 (sub)classes nodes.

4. Results and discussion

Comparing roughly is not sufficient for content-related estimation of visualization maps. Fractal dimension FD provide qualitative evaluation of distributions presented on Figure 2. Calculated values of fractal parameters: self-similarity and lacunarity are attached in Table 1.

Table 1. Fractal characteristics of structures of the visualization maps.

	Map dataset								
	I			II			III		
	0.6: 0.4	0.5: 0.5	0.7: 0.3	1988	1998	2007	W/out class I	Rand	Classes.
Fractal dim. FD	2.761	2.71	2.736	2.86	2.725	2.761	2.731	2.775	2.51
Lacunarity	0.158	0.179	0.201	0.04	0.113	0.158	0.235	0.13	1,03

For convenience calculation results are grouped in comparative series. First ones consist of three different ratio of main to additional classifications. The higher weight of the main classification (the highest tested is 0.7) the more intensive concentration of documents nodes around their main classes and the more clear division into thematic categories. In the contrary case – main classification weight is comparable with additional (0.5:0.5), the categories more merge with one another. The suitable pattern is to be more blur (Picture 1a) thus any structure is disappeared and as a result fractal dimension is the lowest than in two another cases – 2.71. Output visualization map used in all stages of analysis is characterized by ratio 0.6:0.4. The highest FD value (2.761) confirms the fractal structure is the most distinct on this map. For further tests the classification was modified by eliminating the most spacious class I. Less density and smaller FD value (2.731) identifies this distribution.

Lacunarity is a degree of holes distribution and has the lowest value for indeterminate structure. Interpretation of this is the following: the more even distribution of documents nodes the better space filling and less holes is observed. Therefore the perfectly homogeneous localization of objects must

[2] Desaturation is removal from the image the information about colours.

have the smallest lacunarity. To verify this approach random distribution was generated using the same number of nodes (Figure 2d and series III in Table 1). The results prove this assumption is correct.

To interpret a big FD value for random distribution we need to relate to fractal dimension range for organized objects. Trees with linear hierarchy are described by fractal with dimension value above 2 (topological dimension of line is 1, of rectangle is 2). If analyzed visualization map includes some hierarchy structure its fractal dimension must be within a range between 2 and 2.775. The last one is FD value for slight structure i.e. random homogeneous distribution.

In second series fractals parameters of three maps produced for different publishing years are compared (Table 1), called by Garfield longitudinal maps (1994, 1998). In 1988, when ACM classification was in early stage of development, the distribution resembles random sample; FD parameter is very high and equals 2.86. No right structure was found (Figure 2b). The next map shows 10 years later thematic map is more clear. In 2007 the structure of multilevel hierarchy became certainly very definite (last map on Figure 2b); fractal dimension converges to the lower value (2.761).

Original maps are colored by 11 colors. For fractal analysis needs the pictures were desaturated and some information about hierarchy levels was lost. Therefore additionally for longitudinal colored maps spectral analysis steps were performed.

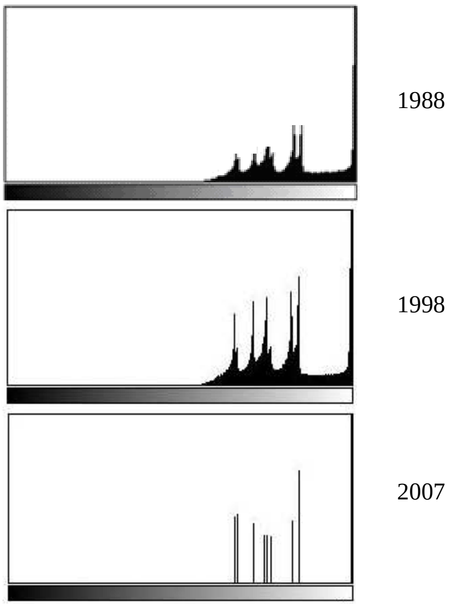

Figure 3. Spectral histograms of classification visualization maps made for different publishing years.

According spectral histograms of visualized collection of scientific articles (Figure 3) it is possible to come to the same conclusion about thematic categorization of ACM classification improved in last decade. Sharp spikes without noise background (which exists on the first two graphs) mean pure colors on visualization maps. Thus the visualization of documents published in 2007 points to a clearly defined organization of thematic categories.

6. Conclusion

Visualization of classification universe depicts scientific knowledge organization in selected domain. Outcome graphic layout facilitates human interaction for exploration and understanding the large amount of data and their correlations. The main problem of visualization techniques there is still no defined quantitative methods to evaluate their results.

In current paper the advantages of fractal analysis for visualization maps are presented. On the assumption that map structure camouflages some fractal its characteristics like dimension and lacunarity are essential for discovering and further insight the data organization. For example hierarchical trees structures as linear formations have fractal dimension between 1 and 2. The higher FD value the distribution is nearer to even type while the information about hierarchy levels is lost. Knowledge about fractal parameters results in choice of optimal visualization as well as finding the stages of Computer Science domain development.

Some researchers with multidisciplinary background study a potential parallelism between fractal theory and knowledge organization (Barát 2009, Scharnhorst 2003, Crowley 2002), that achieves the interdisciplinary perspective of complex structures research. Physics laws are common and universal in the natural world. There are innumerable examples of natural objects can be approximated to fractals. Representation of different concepts been created in human brain as a result of observing the nature must have fractal structure.

References

Barát Á.H.. (2009). The Structures of Concept And its Connection to Sciences. In: *Proceedings of IX ISKO Congres Spain Group, New Perspectives for the organization and dissemination of knowledge*. Valencia: UPV, pp. 372-379;

Börner, K. et al. (2003). Visualizing Knowledge Domains. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology 5, pp. 179 255.

Chen Ch. (2006). *Information Visualization: Beyond the Horizon*. London: Springer, 2nd edition, pp. 143-170.

Crowley Ch. (2002). *Overview of Complexity*. Available online at URL: http://wynchar.com/charlie/Complexity/ overviewOfComplexity.html

Garfield, E. (1994). Scientography: Mapping the tracks of science. Current Contents: Social & Behavioural Sciences, 7(45), pp. 5-10.

Garfield, E. *Essays/Papers on "Mapping the World of Science"*. since 1998. Available online at URL: http://garfield.library.upenn.edu/mapping/mapping.html.

List of fractals by Hausdorff dimension. In: *Wikipedia. The Free Encyclopedia* [on-line]. Available online at URL: http://en.wikipedia.org/wiki/List_of_fractals_by_Hausdorff_dimension

Mandelbrot, B.B. (1982). *The Fractal Geometry of Nature*. W.H. Freeman and Company, pp. 6-20.

Osinska, V. & Bala, P. (2008). Classification Visualization across Mapping on a Sphere. In: *New trends of multimedia and Network Information Systems*. Amsterdam: IOS Press, pp. 95-107.

Osinska, V. & Bala, P. (2010). New Methods for Visualization and Improvement of Classification Schemes – the case of Computer Science. *Knowledge Organization*, 37(2).

Plotnick R.E., Gardner R.H. (1993). Lacunarity indices as measures of landscape texture. *Landscape Ecology*, Vol. 8, nr 3, pp. 201-211.

Scharnhorst, A. (2003). Complex Networks and the Web: Insights Nonlinear Physics. *JCMC* 8 (4). Available online at URL: http://jcmc.indiana.edu/vol8/issue4/scharnhorst.html

Sperber D. (2003). Why Rethink Interdisciplinarity? *Interdisciplines* Available online at URL: http://www.interdisciplines.org/ interdisciplinarity/papers/1.html

Ware, C. (2004). *Information Visualization: Perception for Design*. San Francisko: Morgan Kaufmann, 2nd edition, pp. 43-80.