IX CONGRESO ISKO CAPÍTULO ESPAÑOL (Valencia, marzo 2009)

NONLINEAR APPROACH IN CLASSIFICATION VISUALIZATION AND EVALUATION

Veslava Osińska
Institute of Information Science and Book Studies, Nicolas Copernicus University, Toruń, Poland
e-mail: wieo@umk.pl
Piotr Bała
Institute of Mathematics and Computer Science, Nicolas Copernicus University, Toruń, Poland
e-mail: bala@mat.umk.pl

ABSTRACT:

In this paper we have proposed the novel methodology to visualize classification scheme in informatics domain. We have mapped a documents collection of ACM (Association for Computing Machinery) Digital Library to a sphere surface. Two main stages of visualization processes complement one another: classification and clusterization. Primarily classified documents were visualized and their further clusterization by means of keywords was crucial in evaluation process. For clusters analysis of given visualization maps nonlinear digital filtering techniques were applied. The clusters of keywords were characterized by a local accuracy. Obtained semantic map was included to validation process.

KEYWORDS:

Information visualization, Infovis, 3D visualization, classification tree, ACM, science mapping, semantic map

INTRODUCTION

Information Visualization (InfoVis) is generally defined as a field of research of visual display of information, non-graphical data or knowledge. Making abstract numbers visible have an application in bar plots and pie charts. Instead to read a columns and rows of numbers it is possible to see trends and relationships immediately. Visual representations as a way of communicate ideas have a long history in human activity: from cave paintings to the maps and textual concepts in visual formats. Such disciplines as cartography, labels, sign and wiring design, Computer Aided Design (ACD) and computer graphics have emerged towards the knowledge representation needs.

Two approaches are considered in visualization design. The first one is analysis-oriented practiced by the computer scientists and the other - more artistic is important for computer graphics specialists. The proper synonym of Infovis – infographics has responded to the contemporary movie and advertising needs. Today information graphics surround us in the media, in published works both pedestrian and scientific, in road signs and manuals.

Primarily Visualization was developed as methodology "which employs the largely independent, but converging fields, of computer graphics, image processing, computer vision, computer aided design, signal processing and user interface studies". As a relatively young science branch, it requires underlying universal basis (Kosara, 2007) to allow translation of research results to other disciplines fields. As emphasized many a time the leader in Infovis research professor Chaomei Chen (2006), we lack a theory of visualization – some scientists it call "foundational problem/s of visualization" (Johnson, 2004). Interdisciplinary nature of visualization points to that theory could be comprised of two distinct aspects: one depends only on the underlying data, while the other concentrates on the human response to imagery. The one from the list of 10 top visualization problems (Johnson, 2004) is reliable evaluation of the proposed methods and quantification of the effectiveness of given techniques.

The common factors which decides about the user's respond on visualization image are human perception and cognition. According Kosara (2007) who involved such aesthetic criterion as sublime, which can be considered as something inspiring emotional reaction of observer. Good examples could be art works. The author an artistic and pragmatic aspects put on opposite ends of the sublimity scale what means while the classical technical information visualization is entirely non-sublime, artistic visualizations conduct a high sublime.

Some measures of visualization are defined more precisely, for example visual efficiency or interaction level. Colin Ware (2004) visual efficiency provides the best match of screen pixel to brain pixels. In modeling human visual activity they have calculated an optimal screen size of displaying image. Interaction can be evaluated by the rate of information uptake by the user (Cutrell, 2000), where generally delay time from display to switch attention is measured.

VISUALIZING HIERARCHIES

Hierarchical trees are commonly used for representing hierarchical structures of information, which is the most popular type among other organizational structures like linear, net etc. Hierarchies are presented in file systems, classification schemes, biological classifications of all animals, genealogy, object-programming languages classes diagrams etc. A commonly used strategy is to simplify a network by extracting a tree structure and further to apply a proper visualization technique.

There are many efficient visualization algorithms for hierarchical structures. Well known a classic technique – treemap (Scheiderman, 2006) – utilizes a space-filling algorithm that fills recursively divided rectangle (or circular like in SunBurst application[1]) areas with the components of hierarchy. Tree map have been adopted in such domains as file directory structure, demography, sport statistics or stock prices.

Another popular space-filling approach is self-organizing maps (SOM), discovered by Teuvo Kohonen (Boyack, 2005). SOM refers to the unsupervised learning and artificial neural learning. This technique is used for images or documents of similar topics clusterization. A cluster-based visualization can be useful for many purposes such as getting an overview of documents collection's content.

The problem of displaying a complex information in a limited complex viewing area can be solved by focus+context approach (CAVA, 2001), which helps presenting information about an item in the both information and context spaces. This method provides a visual representation of the entire information space as well as a detailed view of some selected item. These techniques require interaction mechanisms to change the focus, usually showed in detail, keeping the context as stable as possible. The browsers apply this method are called hyperbolic.

The previous chapter exposes the common problems regarding visualization results. In most papers about visualization an accuracy of given visualization layout is deduced. There has been a little written on how to quantitatively evaluate the accuracy of relatedness measures or the resulting maps.Generally the outcome objects layout is evaluated arbitrary with participation of domain experts, visualization results are no quantificated. A few works report about methods to calculate quality measures applying in final stage of visualization process. The authors (Samoylenko, 2006) propose a new framework for assessing the performance of relatedness measures and visualization algorithms that contains four factors: accuracy, coverage, scalability, and robustness. Relatedness measures are used for many different tasks such as generating of maps, or visual pictures, showing the relationship between all items from these data.
The validation of whole science map in (Boyack, 2005) used the Institute of Scientific Information (ISI) journal classifications. The correspondence of visual clusters to ISI category assignments determines the validity of proposed earlier eight different similarity measures. To calculate a quality of clusters assignments was done by using Shannon's formula for entropy.

In current work we have visualized primarily classified documents and their further clusterization. Automatic classifying of documents occurs with top-down scheme which starts from categories (classes) and then assigns items to a given categories. The opposing process - clustering is characterized by a bottom-top approach. Earlier solutions of classified objects 3D Visualization were based on the hierarchical structure 3D space where the root node is to be located in the centre and all sub-nodes will spread out in all directions around of the central nodes.

Methodology

Our work was concentrated in visualizing of classified documents and further constructing a new graphical representation of original classification scheme. Experiment's data were collected from the ACM (*Association for Computing Machinery*) digital library[2] which originally were classified into classes and subclasses.

[1] www.gvu.gatech.edu/ii/sunburst

2 http://www.acm.org/dl

The main task was concerning the similarity metrics of documents. The space of primary classification tree can not be used for similarity measure because of their linearity. Data population was the highest on the lowest levels for the most classes. If some sublevels nodes split conceptually the documents were appeared in both (sub)classes. We assumed that the topic similarity between classes is proportional to the number of recurrent documents. As closer thematically two subclasses the more common articles they include. This pair of classes must be crossed in typical dendrogram tree. And inversely dissimilar subclasses contain no common data. Count and normalize the number of common documents for every pair of classes and subclasses it is possible to construct matrix similarity. Dimension of square matrix is equal the number of all occurred in the data collection classes and subclasses.

As a target information space we have chosen the sphere surface because of their symmetric and curved surface maps the distances between the data more effectively than a plane. Furthermore sphere is comfortable in navigation and retrieval processes. We used multidimensional scaling algorithm (MDS) to reduce matrix dimension to three in order to transfer the representation model into Euclidean space. MDS is statistical techniques often used in information visualization for exploring similarities or dissimilarities in data. To build an optimal representation, the MDS algorithm minimizes a criterion called Stress. The smaller the stress value, the better is the fit of the reproduced distance matrix to the input distance matrix.

On the basis of given coordinates we have constructed a classification sphere. This graphical 3D representation allowed to visualize such attributes of classes as classification code, quantity, tree level, proximity by means of colour, size, position and transparency degree respectively. We obtained a multidimensional navigation space where the relevant information can be conveyed in a compact display, including topics, relationships among topics, frequency of occurrence, importance and evolution.
Next we have analyzed a map of documents clusters by means of nonlinear digital filtering techniques.

VISUALIZATION PROCESS

DATASET

An experiment's details, used data and data processing stages were described in our previous work (Osinska, 2008). We have collected the abstracts of publication from ACM Classification of Computing System (CCS) digital library. Besides index codes the main metadata as a title, keywords and general terms have been extracted. The full classification scheme involves three coded level tree.[3]. The upper level consists of 11 main classes which are listed on Figure 1. Every category name start with a suitable capital letter: from A to K. CCS is still updating and therefore a new subdivisions are appeared with “New” label or some of existing categories are prepared to removing with suitable label “Reviced”.

In selection the CCS literature collection we have motivated by two reasons. The authors are most familiar with what research domain describes their work. In addition digital library system provides on-line access to all classified abstracts of publications.

[3] http://www.acm.org/class/1998/

A. General Literature
B. Hardware
C. Computer Systems Organization
D. Software
E. Data
F. Theory of Computation
G. Mathematics of Computing
H. Information Systems
I. Computing Methodologies

Figure 1. - ACM CCS main classes

After we rejected duplicates (the same documents appear in different classes and different levels) the total number of documents became 37 543. The first objective of current work was focused to spread out this quantity of nodes on a sphere surface most efficiently. The final number of classes codes including all levels and two sublevels was equal 353. Similarity matrix consist of 353 rows and 353 columns. MDS algorithm was used to reduce matrix dimension to three. Finally all collected documents were mapped on a sphere surface with preserving their features.

MAP LAYOUT

Every of 11 main class was marked by different color. A set of effective colors for coding was chosen according perceptual factors (Ware, 2004): red, green, yellow, blue, pink, cyan, gray, orange, brown, black, purple. The subclasses are described by appropriate color lightness. Therefore color palette has been extended to the number of colors be attributed to three levels nodes: 11×3=33. On the Figure 2 we can see the visualization of 347 classification nodes mapped to the classification surface.

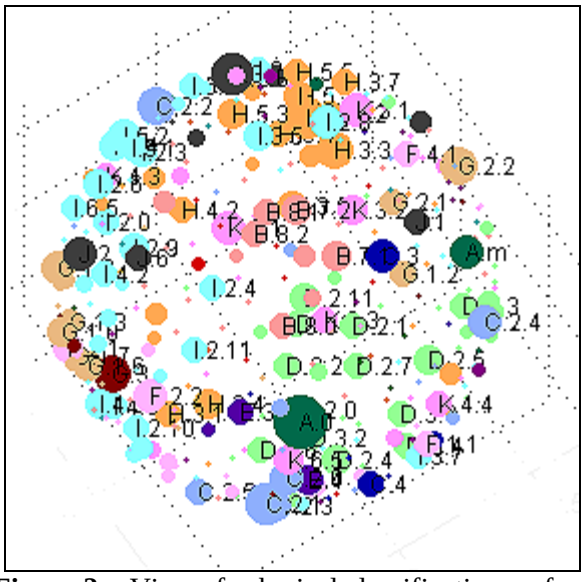


Figure 2. - View of spherical classification surface

Every dataset document is characterized by the main class code and a several codes of additional classes. These classes nodes form such figures as: lines, triangles or polygons depending of their quantity. The centroid of such figure determines a document's location on a sphere surface. Essential task was to establish the weights main and additional classifications. As higher weight of additional classification the document positions farther from a primary class node. For distribution

estimation we have selected three values of relations: 0.7:0.3, 0.6:0.4 and 0.5:0.5. We have noticed the clusters were disappeared for primary class weight 0.5 and lower. Distinct clusters in the map allow to validate documents assignment to the original categories. From the opposite side minimize a weigh of additional classes causes the objects gather around of primary classification node. Observed dependency allow to choose primary and additional classification weights 0.6:0.4 respectively. All further calculations and simulations were performed for the weights 06:04.

Figure 3a) demonstrates the data collection layout on the sphere surface. The documents inherit the color of the main class. For easer processing and analysis of given map we used their cartographic projection on a plane (Figure 3b). In this map longitude and latitude values were transformed as x and y coordinates.

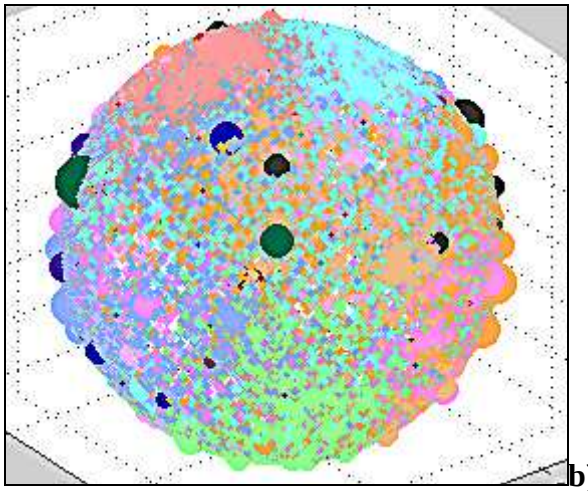
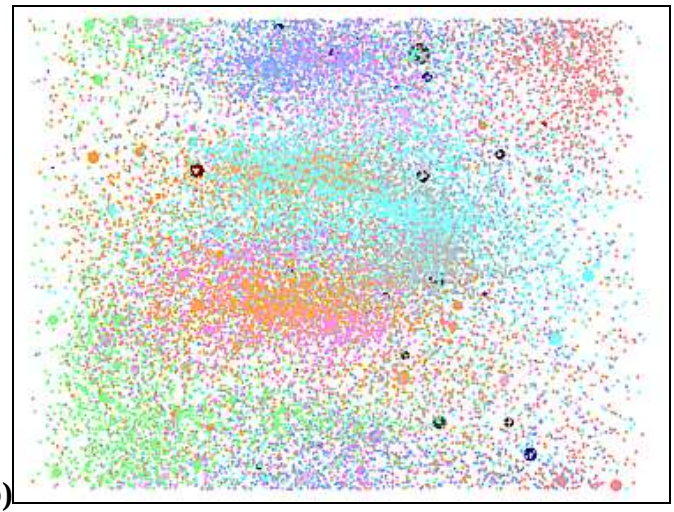
a)  **b)** 

Figure 3. – a) All classes documents set on a sphere; **b)** cartographic view of sphere surface

DATA ANALYSIS

We got a map of all 37 543 objects been clustered into one of 11 colors patches. As a closer inspection shows color dots uniformly fill a sphere surface and form color patches according main classes themes. The clusters were characterized by dissolved border and the next steps of data processing were required.

In order to do a better analysis of clusters borders we applied image processing algorithms. The median filter is a non-linear digital filtering technique, often used to remove noise from images or other signals. Relating to received map we can treat as noise some single, distant points which disturb final pattern of clusters. First the median filter and next contour filter for edge detection were applied. Resulting image besides distinct clusterization demonstrates (Figure 4) such properties as splitting and crossing of clusters. In fact a clusters are mapping categories of (sub)classes, so a places of their overlapping detected through colors mixing determine similar themes of original classification tree. On Figure 4a the mostly uniform distribution of documents objects is displayed. Picture 4b shows, that the ontologically different Hardware (B class) and Software (D class) are distributed in the opposite corners (poles in case of sphere). However class C as networks category places between them because of both problems are represented. In the aftermath one can count a number of clusters and validate thematic diversity of main classes.

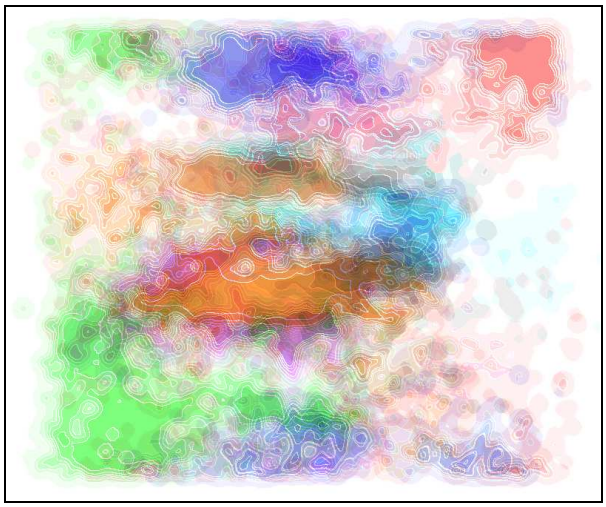
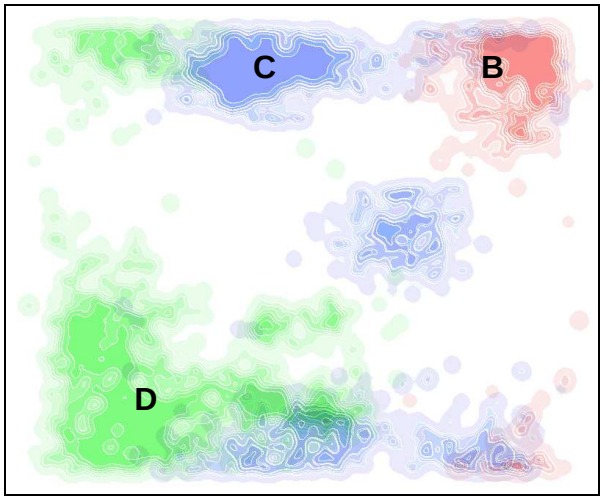
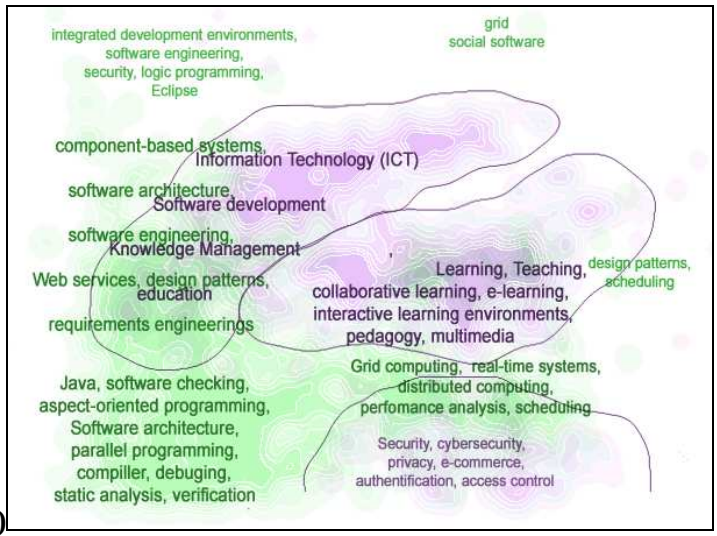
a)  b) 

Figure 4. – a) all and; **b)** BCD classes maps after image processing

Second phase of research was documents identification within a clusters and keywords extraction. Statistically obtained a sets of most frequent keywords of each cluster have been used to caption them. Analyzing the results it was important to consider any keyword depending on other terms belonging to one cluster. Clusters were captioned using the color of proper main class as one can see on Figure 5. Most important feature was that different colors but close lying terms show to similar topics. In this case information about position is more relevant than information about a color and therefore main class. Apparently semantic map was organized logically. The clusters layout was characterized by a local accuracy. On the poles such topic as security, privacy, authentification and cryptography have been concentrated one another. Information Systems keywords: Visualization, clustering, Web services, Knowledge Management are in the close neighboring terms belonging to different class: ontologies, multi-agent, mashine learning, grid computing, distributed computing (see Figure 5b). Clusters are designated by teaching and education topics: e-learning, pedagogy, multimedia have been located near to Software development, OO languages and Information Technology, as demonstrates Figure 5a.

a) 

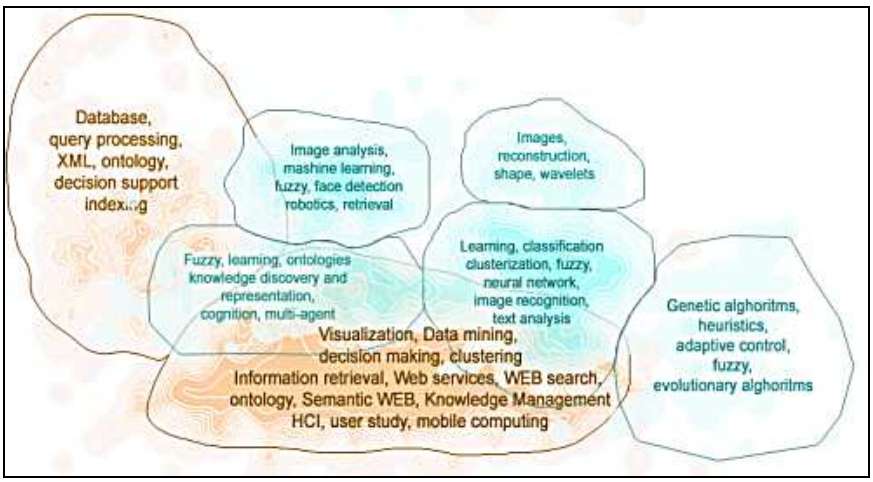
b) 

Figure 5. – Semantic maps of keywords for classes **a)** D, K and **b)** I, H

CONCLUSION

In this paper we have proposed the novel methodology to visualize classification scheme in informatics domain. A sphere surface was chosen as mapping and navigating geometry. Documents were extracted from ACM (Association for Computing Machinery) Digital Library. We have applied both class similarity metrics and MDS technique to the collection of articles abstracts published in 2007 year. Dataset counts 37543 items and the number of received classes nodes achieved 353. To overcome the incorrectness of linear measures in indexes distances we calculated similarity matrix of themes and MDS coordinates. Uniform distribution of all documents on a mapping surface provided that this is proper strategy of examined classification trees visualization. Mapping subject classification scheme into a sphere gives more possibilities for matching the distances between the classes than in classical hierarchical tree case. Proposed method to visualize classification scheme is suitable to reach nonlinearity in subjects content visualization.

Matlab environment where visualization model was performed provided easy browsing, rotating and zooming with relation to whole surface of classification sphere. It was possible to monitor topology of class nodes from any perspective. Interface provided also the representation of documents set belong to selected class. The concept of this visualization is convergent with the following leitmotiv: "The eye seeks to compare similar things, to examine them from several angles, to shift perspective in order to view how the parts of a whole fit together".

For analysis of given visualization maps nonlinear digital filtering techniques were applied. The median filter removed noise, and the edge detection fractal – based algorithm gave us final information about different main classes frontiers. Final clusters were designated by the keywords extracted from statistical calculations. Semantic map of keywords shown revealed such important property as local accuracy which was included to validation process. Accuracy at this point means similarity in both paradigmatic and intuitive comprehension of themes.

Next research will be oriented towards keywords map detailed analysis and further its confrontation with existing classification scheme. We plan also to add one dimension more - a time and to repeat experiment for the wider time range. Sphere model of data visualization and simulation could facilitate the studies of Computer Science classification evolution. ACM offers one of the biggest digital collection of literature in Computer Science domain. This allows us to observe dynamics in Information Technology and Engineering development.

By mapping a classification we constructed the dynamic knowledge space. Appropriate application with animated layouts may demonstrate domain history and prediction what subfield is far-reaching what is decayed. Resulting map should be flexible both to visualize full coverage of a new categories intellectual content and to reduce thematic space old categories case.

For a scientists working in different research areas this visualization will be very useful. Especially on a interdisciplinary field the scientists can predict the branch growing dynamics. This visualization method in a many time cycles give us a chance to simulate a future structure of proper knowledge.

BIBLIOGRAPHY

BOYACK, Kevin W., KLAVANS Richard and BÖRNER, Katy et al. Mapping the backbone of Science. In: *Scientometrics.* 2005. Vol. 64, No. 3.

CAVA, Ricardo Andrade, FREITAS, Dal Sasso. Visualizing Hierarchies Using a Modified Focus + Context Technique. In: *Proceedings of IEEE Information Visualization (Interactive Poster).* 2001.

CHEN, Chaomei. *Information Visualization: Beyond the Horizon.* London: Springer, 2nd edition, 2006. ISBN-13: 978-1846283406.

CUTRELL, E.B., CZERWINSKI, M. and HORVITZ, E. Effects of instant messaging interruptions on computing tasks. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems, Extended Abstracts* .ACM Press, New York, 2000.

ELERT, Glenn.*The chaos hypertextbook.* [online]. 2007. [access: 22 Nov 2008]. Accessible on Web: <http://hypertextbook.com/chaos/>

JOHNSON, Christopher R.. Top Scientific Visualization Research Problems. [online]. In: Visualization Viewpoints. IEEE Computer Society. 2004. [access: 17 Nov. 2008]. Accessible on Web: <http://erie.nlm.nih.gov/evc/meetings/vrc2004/position_papers/johnson.pdf>.

KOSARA, Robert. Visualization Criticism – The Missing Link Between Information Visualization and Art. In: *Proceedings of the 11th International Conference Information Visualization.* 2007, pp. 631-636.

OSINSKA, Veslava, BALA, Piotr. Classification Visualization across mapping on a Sphere. In: *New trends of multimedia and Network Information Systems*. Amsterdam: IOS press. 2008. ISBN 13: 978-1-58603904-2.

SAMOYLENKO, I., CHAO, T.-C., LIU, W.-C. and CHEN, C.-M. Visualizing the scientific world and its evolution. In: *Journal of the American Society for Information Science and Technology*. 2006. Vol. 57, Issue 11, pp. 1461-1469.

SCNEIDERMAN, Ben. *Treemaps for Space Constrained Visualization of Hierarchies*. [online]. Report. 2006. [access 15 Nov. 2008]. Accessible on Web: http://www.cs.umd.edu/hcil/

WARE, Colin. *Information Visualization: Perception for Design.* San Francisko: Morgan Kaufmann, 2nd edition, 2004. ISBN-13: 978-1558608191.