Visual Analysis of Classification Scheme

Veslava Osińska

Institute of Information Science and Book Studies, Nicolas Copernicus University,
ul. Gagarina 13a, 87-100 Toruń, Poland <wieo@umk.pl>

Veslava Osińska received an MSc in physics from Vilnius University and a Ph.D. in Information Science and Bibliography from Nicolaus Copernicus University in Torun (Poland), where she teaches Information and Communication Technology and Computer Graphics. She has applied her Computer Science background and programming skills to research areas which include effective visualization of multidimensional information, as for example, bibliographical data generated in digital libraries. She is a member of the Polish Chapter of the International Society for Knowledge Organization and Polish Computer Science Society.

Osińska, Veslava. **Visual Analysis of Classification Scheme.** *Knowledge Organization, 37(4)*, 299-306. 25 references.

ABSTRACT: This paper proposes a novel methodology to visualize a classification scheme. It is demonstrated with the Association for Computing Machinery (ACM) Computing Classification System (CCS). The collection derived from the ACM digital library, containing 37,543 documents classified by CCS. The assigned classes, subject descriptors, and keywords were processed in a dataset to produce a graphical representation of the documents. The general conception is based on the similarity of co-classes (themes) proportional to the number of common publications. The final number of all possible classes and subclasses in the collection was 353 and therefore the similarity matrix of co-classes had the same dimension. A spherical surface was chosen as the target information space. Classes and documents' node locations on the sphere were obtained by means of Multidimensional Scaling coordinates. By representing the surface on a plane like a map projection, it is possible to analyze the visualization layout. The graphical patterns were organized in some colour clusters. For evaluation of given visualization maps, graphics filtering was applied. This proposed method can be very useful in interdisciplinary research fields. It allows for a great amount of heterogeneous information to be conveyed in a compact display, including topics, relationships among topics, frequency of occurrence, importance and changes of these properties over time.

1.0 Introduction

With the exponential growth of Internet resources, it has become more and more difficult to find relevant information from one hand and organize professional information services from the other. From this perspective, this article will focus on the visual analysis and evaluation of a classification system in Computer Science (CS) which has evolved into a very dynamic domain. New computer technology branches emerge, some of them split into smaller ones, while other subfields of the CS domain have disappeared. Professional CS classifications are nowadays challenged by rapid changes in taxonomy and users needs to retrieve relevant information. To strengthen research in this area there is a need to build upon innovative efforts of information visualization (Infovis), computer, and library scientists.

While library resources are continuously extended, LIS researchers may use new tools derived from Infovis methodologies in order to support collection management. Visualization of the complex structures of large amounts of information may help in understanding relations between components and visually searching relevant information. Thus visualization became a phase of data analysis. In the last decades, Infovis projects have specialized in both LIS and medical data representation tasks (Börner 2003, Chen 2006, Kosara and Miksch 2002). The main aim of the presented work is to visualize the chosen classification scheme and its universe. In our opinion, only one publication presents an effort to visualize classification. For example, treemap visualization of a specific library collection is performed to facilitate document retrieval in bibliographic collections (Pfeffer et al. 2008).

First, modern techniques of a hierarchy mapping are introduced. Some of them inspired the author to make the final conception of the information space. The results encouraged researchers to find new experimental problems and the ways of solving them. The primary task – visualization – developed further into several other tasks with more specific interests such as: classification scheme evaluation, domain evolution, and documents retrieval. Results of the experiments show that librarians may use proposed methods in classification modernization, evaluation, and analyzing, as well as in studying the scientific domain organization.

2.0 Hierarchical representation of the structures

2.1 Trees

A natural way to present the hierarchical nature of data structure is a tree. The starting element, root, is usually positioned on top. The names of relationships between nodes are modelled after kinship relations. A node is a parent of another node if it is one level higher than subordinate nodes, children. Sibling nodes share the same parent node. Tree diagrams impose linear order, vertical direction (Figure 1a). In a tree structure, information disseminates one way: from parents to children and vice versa. Hierarchical information is the most frequent type of data occurring in the human environment. Such hierarchy exists in library classification systems, genotype systems, genealogy data, as well as computer directory structures and object-oriented programming languages class definitions. If paths between siblings became available, a tree structure evolves in the net. E-book texts with hyperlinks of chapters can be an example of such type of information space.

Traditional library classifications are presented in deductive, top-down schemes with a set of mutually exclusive classes (Jacob 2004). Exclusivity means that a given entity must be assigned to one and only one class within a system of mutually exclusive and non-overlapping classes. The top class is the most inclusive class and depicts the domain of the classification. Being a system of classes and subclasses, a classification is organized according to predetermined and essential properties of a set of entities. Construction of the scheme involves the logical process of division and subdivision of the original universe. In consequence, the hierarchical tree of generic relationships is formed. Within superordinate classes, more or less subordinate classes are nested. To simplify the task of classification visualization, it is convenient to limit it to its monohierarchical structure; this is the case with a classification universe that encompasses only one hierarchy tree.

Kwasnik (1999) described the browsing of a classification scheme in the following way: "[it] involves moving down the hierarchy, from superordinate to subordinate and from left to right, to generate a series of relationships between classes that can be translated into the linear order of the library shelf." This feature of linearity, amongst exclusivity, aggregation, and infinite hospitality, is identified as a characteristic of a bibliographic classification scheme (Shera 1965). Librarians appreciate a tree as a way of representation of the relative placement of the entities because of its good local visibility (child nodes frequency). On the other hand, some disadvantages are associated with this form of knowledge representation that are specific to library classifications (Kwasnik 1999):

1) Lack of flexibility in adding new entities and coping with new knowledge emergence. This often requires changing the general shape of the tree, which is determined a priori;
2) Partial inference: trees are limited in the representation of knowledge volume;
3) Only vertical direction of information dissemination, therefore imposing the same understanding on individuals;
4) Selective perspective: by emphasizing a certain relationship, a tree masks other equally interesting relationships.

The distinct methods to display hierarchy structures will be discussed below. Tree mapping inspired many Infovis researchers, HCI (Human-Computer Interaction) experts and strictly commercial data mining applications designers (Börner 2002).

2.2 TreeMaps

The main distinguishing feature of a treemap technology relies on unlimited recursive construction of nested geometric primitives: rectangles, circles, arcs and so forth – thus mosaic plots can be created. This property allows a final layout to be extended to hierarchical data with any number of levels. This idea was invented by Ben Shneiderman (1998-2009) in the early 1990s "in response to the common problem of a filled hard disk.… Since the 80 Megabyte hard disk in the HCIL was shared by 14 users it was difficult to determine how and where space was used.

Finding large files that could be deleted, or even determining which users consumed the largest shares of disk space were difficult tasks." According to treemap algorithms, one must divide an original rectangle (or another shape) space into sub-rectangles as many times as number of levels in the structure. This technique sometimes is called tiling or puzzling. Sub-rectangles have an area proportional to a specified dimension of data, usually size or population of nodes (Figure 1b). Colour is used to separate a type of data (for example electronic files format in directories trees). While the colour and size dimensions are correlated in this way with the tree structure, the occurred pattern can reveal interesting properties of data. A second advantage of treemaps is an efficient use of space: it is possible to legibly display thousands of items on the screen simultaneously. Treemaps initiated an entire software generation serving for visualization of large datasets. Some of them are open source or demo versions such as the Treemap 4.0 tool designed by the HCIL at the University of Maryland (HCIL 2003).

If one replaces rectangles by circles, the circle treemap will arise (Figure 1c Sunburst mapping which also

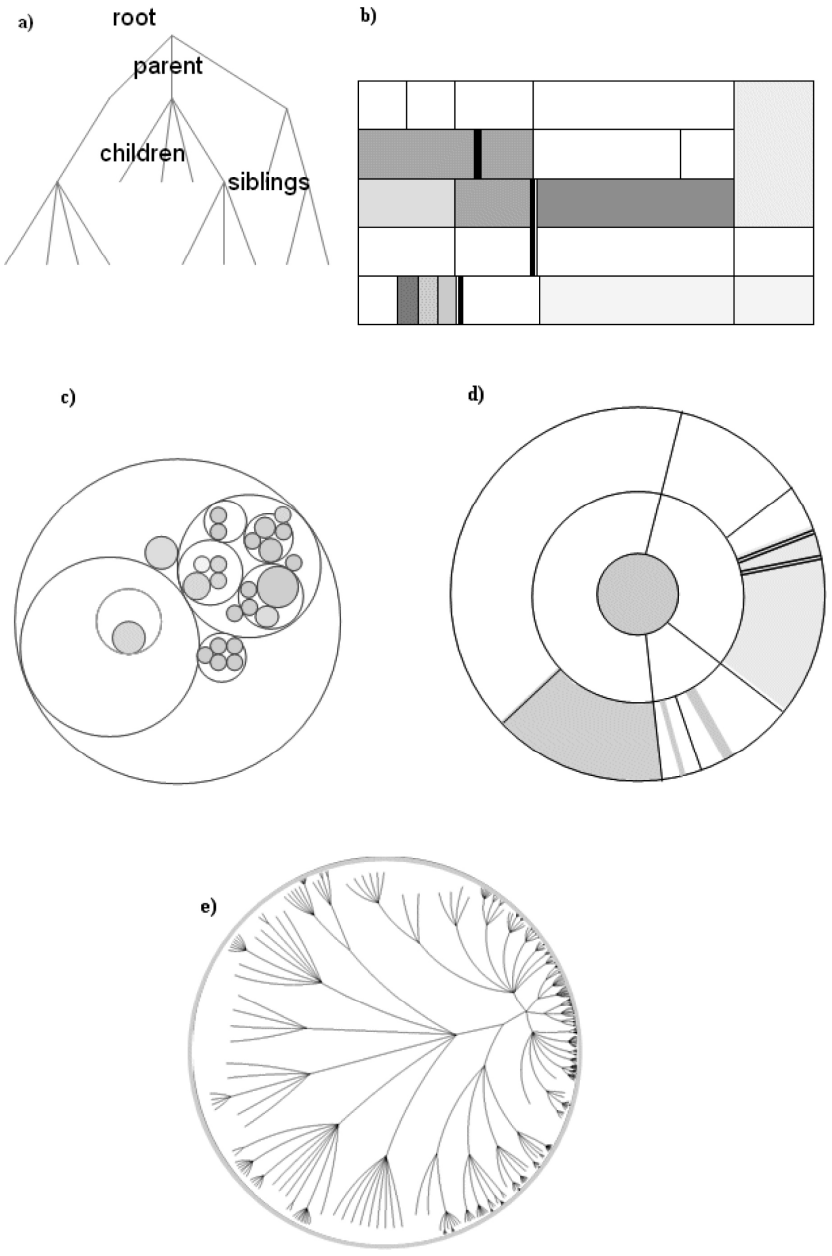

Figure 1 Graphical representations of hierarchies: *a)* traditional – as tree; *b)* rectangle treemap; *c)* circle treemap; *d*) sunburst map; *e)* in the hyperbolic space.

exploits circles, is shown in Figure 1d. The root node is located at the center, each successive level drawing farther out from center. An infinity of hierarchy levels can be represented. Size and a type of data are identified by the sweep angle and colour of the item, respectively.

An important technique to magnify an exploration space is workspace construction in hyperbolic 3D geometry. The first applications that used "fisheye technique" (called also "focus+ context") are hyperbolic browsers – this approach assures more place to visualize the hierarchy (Chen 2006). This leads to the convenient property that the circumference of a circle grows exponentially with its radius, which means that exponentially more space is available with increasing distance (Lamping et al. 1995). It is possible to study the hyperbolic view of various types of data and construct a graph upon samples using online available applications such as], Hyperbolic 3D (Munzner 1998) or Walrus (CAIDA 2005-2009). Figure 1e illustrates such an hyperbolical treemap.

3D visualization is very promising because of the continuously growing potential of hardware. Due to current users' requirements, 3D models are standard in any type of computer games, simulation, and movies. The above mentioned shapes and properties of treemaps, especially sphere determination of an information space, were used in the conception of the classification visualization being presented.

3.0 Visualized Classification

The tested classification, Computing Classification System (ACM 1998), is a subject classification for computer science devised by the Association for Computing Machinery, the first scientific and educational computing society in the world. The last version of CCS was published in 1998 and is still being updated.. The ACM digital library is a vast collection of citations and full text (accessible for members) from ACM journals, newsletter articles, and conference proceedings. Citations consist of a title, author, publication data, abstracts, references, symbols from CCS, and other metadata.

The ACM CCS consists of a four-level tree [containing three levels coded by 11 capital letters (from A to K) and numbers plus a fourth uncoded level], General Terms, and implicit subject descriptors. Thus, the upper level consists of 11 main classes:

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
J. Computer Applications
K. Computing Milieux

Each top-level category has two standard subcategories: "general", coded with "0", and "miscellaneous", coded with "m". For instance, H.0 denotes the "general" subcategory of Information Systems, while H.m describes its miscellaneous subcategory. CCS is still being updated and therefore new subdivisions appear with the "New" label while some of the existing categories are marked as "Revised". Besides a primary classification, every document may be assigned additional ones; as a result, two or more classification trees will be generated. Thus, for example, the book *Semantic Digital Libraries*, by S.R. Kruk and B. McDaniel, Springer, 2008, will have the following classifications:

Example:

Primary Classification:

H. Information Systems
H.3 INFORMATION STORAGE AND RETRIEVAL
H.3.7 Digital Libraries

Additional Classification:

A. General Literature A.m MISCELLANEOUS	I. Computing Methodologies I.2 ARTIFICIAL INTELLIGENCE I.2.4 Knowledge Representation Formalisms and Methods Subjects: Semantic networks

4.0 Methodology

Our previous articles describe in detail the construction of a new graphical representation of an original classification scheme (Osinska and Bala 2008, 2009). Metadata of articles published in 2007 were com-

pleted and processed using ACM Digital Library. The key feature in the method presented was the exclusiveness of CCS classification. Therefore, overlapping classes and subclasses will appear simultaneously among document's citation attributes (as in the above example). According to the author's assumption, these "common" articles must decide on the semantic similarity of thematic categories of classification. The main idea consisted in estimating of co-occurrences of classes, i.e., counting of common documents for every pair of classes and subclasses. Such similarity of co-classes (themes) is proportional to the number of common publications. The final number of all possible classes and subclasses in the collection was 353. Similarity matrix of co-classes had the same dimension. In order to decrease such a high dimension, MDS (Multidimensional Scaling) 3D plot was used. A sphere surface was chosen as target information space because of its symmetry, curved surface, and ergonomic feature (Osinska and Bala 2009). Classes and document nodes locations were mapped on a sphere by means of Multidimensional Scaling (MDS) coordinates.

5.0 Mapping results

By projecting a sphere surface on a plane according to cartographic rules, it is possible to analyze the visualization layout of classes and items nodes. Moreover, nonlinear digital filtering can be applied to given pattern of 2D maps (see Figure 2a, b). After evaluating the Computer Science domain evolution by means of longitudinal maps, i.e., a series of chronologically sequential maps (Garfield 1994), a novel technique derived from fractal theory was successfully used.

5.1 Classes and documents visualization

Figure 2 presents the resulting visualization layouts on sphere surface (a) and its projection on a plane (b). There are 3 attributes: colour, luminosity of colour and size of node were used to indicate main classes, subclasses level and classes population respectively. The documents (37,543) inherit the colour of the main class, therefore the final patterns shown on Figure 2 consist of 11 colourful irregular spots.

Finally, non-linear graphic filtering techniques were applied to remove noise and detect cluster edges, median and contour filters used sequentially. These algorithms enabled access to essential information about the main classes' frontiers and mutually related fields as well as the study of thematic diversity (clusters of some classes are shown in Figure 3).

5.2 Keywords mapping

In the next stage of the research, such attributes of documents as keywords were used. Within each given cluster, statistical ranking of keywords was performed. Figure 3 illustrates clusters of class I and the keyword sets that characterized them. Analyzing each cluster in this way, it is possible to build a semantic map of all classified documents based on the keyword set.

It is worth comparing the first map of classes' themes with the second (keywords) in respect to semantic conformity. Previous works report about local accuracy of tested maps, that means paradigmatic and intuitive comprehension of themes (Osinska and Bala 2009). This issue will be considered more in detail below, in the Discussion section..

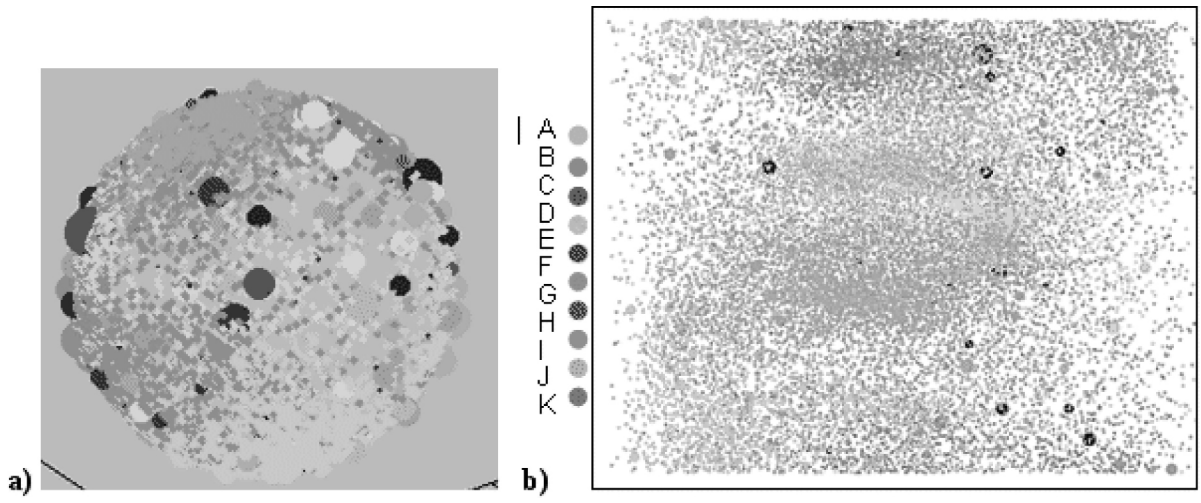

Figure 2. a) Class and document nodes visualization on a sphere surface; *b)* cartographic projection of previous layout

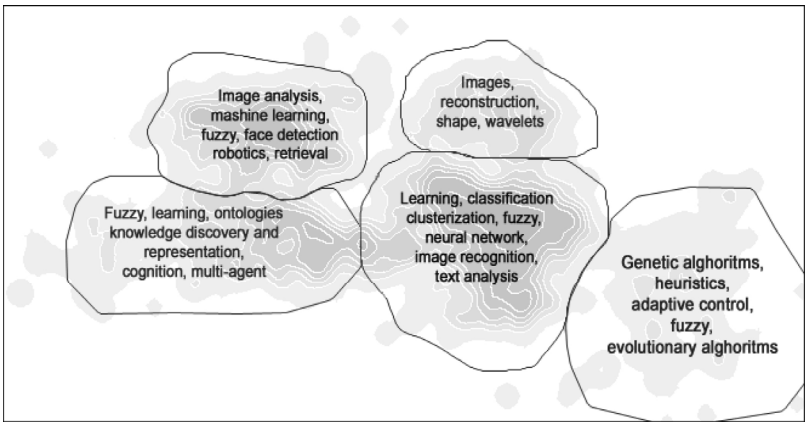

Figure 3. Map of keywords within 5 clusters of main class I.

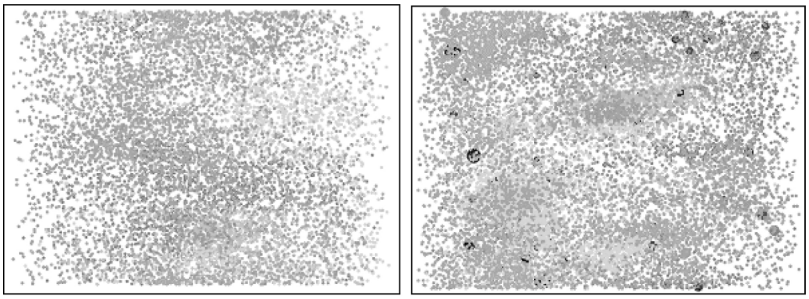

Figure 4. Visualization maps of classified documents from ACM digital library published in: *a)* 1988 *b)* 1998

5.3 Longitudinal Mapping

Using series of chronologically sequential maps, one can study how knowledge advances and knowledge organization change. The term longitudinal mapping was first introduced by Garfield (1994, 1998) to describe this method of domain analysis . He emphasized that longitudinal maps become forecasting tools because main trends can be detected by observing changes from year to year.

A set of visualization maps of ACM documents published every ten years was prepared. This approach should reveal essential changes in Computer Science (CS) literature within the time frame of the Computing Classification System existence. On the basis of these results and computing history expert knowledge the inference about classification evolution, CS integration with other domains and future trends is available. Tested collections amount to 209, 545, 19,950, 27,149, and 37,543 classified documents from 1968, 1978, 1988, 1998, and 2007, respectively. The first two layouts present small quantity of nodes without significant patterns and can be omitted in further analysis.

Information Systems is a category from which the CCS scheme started to grow. Since 1998, arrangement of classes such as B. Software, C. Computer Systems Organization, and H. Information Systems are detectable. It is possible to conclude from three visualization layouts (Figures 1b, 4a,b) that the time when clusterization started relates to the 90s. This means that the two last decades provide a close adaptation of the CCS scheme to the ACM digital library resources.

6.0 Discussion

ACM digital library editors continuously make corrections to the CCS scheme. They are responsible for the timeliness of the updating of the classification tree, aligning it with the dynamics of computer technologies. The authors of articles are well-acquainted with the Computer Science domain, both in practice and theoretical terms. The ACM website provides detailed instructions to authors about how to classify their documents (ACM 2010). They have to add keywords and to describe the documents' main and additional categories as well as apply subject descriptors. ACM editors can correct the classi-

fication assignment and make final decisions about the trees' topology. It should be mentioned that the characteristics of the keywords are an effect of the author's competence and exactness. While the primary visualization maps were based on class correlations, the keywords maps were constructed by means of keywords. The latter became the way to verify the first graphical layout. Therefore, as these independent knowledge paths are confronted, two separate social structures can be identified as a modern approach to domain analysis (Hjørland 2002).

Another important feature of the map resulting from clusterization is the arrangement by colour. With this visualization process the original taxonomy was discarded. All document nodes inherit colour from main classes so that they form clusters which present only one hierarchical level. The reduction of the structure hierarchy from three to one was noted. If the outcome of clusterization reflects the logical categorization of modern Computer Science literature then the CCS scheme will not need so many levels of structure. Consequently, the coverage of thematic-semantic categories within the clusters on the visualization map can inform about the quality of the organization of the input classification.

The formal analogy of the resulting clusterization with faceted classification will be considered below. Faceting classification has been a major development in current library research, especially regarding Information Retrieval tasks (Mills 2004). Analytico-synthetic classification systems are inductive, bottom-up schemes generated through a process of analysis and synthesis (Jacob 2004). A facet classification comprises logically defined, mutually exclusive, and collectively exhaustive aspects, properties or characteristics of a class or specific subject (Taylor 2006). In faceted systems, instead of pre-determined, taxonomic order there are multiple ways of classification information assignment.

The first process in the present research was statistical analysis of co-(sub)classes of initial classification. Clusterization is made on the basis of graphical representation and can be considered as the next step, synthesis of clusters. In traditional faceted classification (Adkisson 2003) analysis provides breaking down subjects into basic concept (semantic analysis) and synthesis – functional categorization. The present case shows the opposite method: while the synthesis is of a semantic nature, the analysis explored the configuration of nonlinear features of the original classification scheme, as the primary units of analysis – (sub)classes symbols – relate to themes and areas of scientific research. The resulting thematico-semantic clusters can be considered as final multi-aspect facets with dynamical parameters such as number of data points, density, size and foremost keyword sets.

Original taxonomical classifications impose a vertical flow of information and thus provide a top-down exploration of the structure. Faceted classifications, used in faceted search systems, enable users to browse data along multiple paths corresponding to different sorting of the facets (Taylor 2006). Similarly, the resulting information space allows the retrieval of similar documents in neighbouring locations, irrespective of navigation directions and primary hierarchy of categories.

7.0 Conclusion

This work presented a novel visualization method of Computing Classification System (CCS) and its classified universe consisting of a large body of scientific literature in the Computer Science domain. An analysis of the initial classification scheme by independent thematic categories was proposed. The basic feature of the original information space transformation into clusters relies in the reduction of the hierarchy. It is noted that one level structure is sufficient to present a logical division of the Computer Science literature in a graphical way. Coverage of thematic-semantic categories within the clusters on the visualization map can report the quality of the organization of input classification. As a result, the local accuracy within the clusters of visualization maps was observed. Citations gathering and data processing were repeated for articles published in the years 1968, 1978, 1988, 1998, and 2007. The longitudinal mapping allows the discovery of the structure of knowledge within the CS domain as well as the social patterns of its scientific output.

With the proposed visualization method, librarians could depict the organization of the contemporary knowledge domain, investigate multidisciplinary fronts of research and predict future trends. The author demonstrated its usefulness in LIS problems such as evaluation of classification schemes and their further improvement. The method can be functional in automatic classification tasks (Golub 2006) as well as, for example, in automatic generation and updating of classification trees. Scientists from interdisciplinary research fields will be able to make full use of the multidimensional navigation space. The approach described allows for large amounts of heterogeneous information and multidimensional data to be conveyed in a compact display as well as for the retrieval

of data by topics, relationships among topics, frequency of occurrence, and relevance and changes of these properties.

References

ACM. 1998. *ACM computing classification system. Association for Computing Machinery.* Available http://www.acm.org/about/class/1998.

ACM. 2010. *How to use the* Computing classification system. Available http://www.acm.org/about/class/how-to-use.

Adkisson, Heidi P. 2005. Use of faceted classification. *Web design practices*. Available www.webdesign practices.com/navigation/facets.html.

Börner, Katy et al. 2003. Visualizing knowledge domains. In Cronin, Blaise, ed., *Annual Review of Information Science & Technology* 5. Medford, NJ: Information Today, pp. 179-255.

Browse maps. Places@Spaces: Mapping Science. Available http://scimaps.org/maps/browse/

CAIDA (2005-2009). *Walrus – Graph visualization tool. The Cooperative Association for Internet Data Analysis*. Cooperative Association for Internet Data Analysis. Available http://www.caida.org/tools/visualization/walrus/

Chen, Chaomei. 2006. *Information visualization: Beyond the horizon*, 2nd ed. London: Springer.

Garfield, Eugene. 1994. Scientography: Mapping the tracks of science. *Current contents: social & behavioural sciences* 7n45: 5-10.

Garfield, Eugene. Since 1998. *Essays / Papers on "Mapping the World of Science".* Available http://garfield.library.upenn.edu/mapping/mapping.html.

Golub, Koraljka. 2006. Automated subject classification of textual web documents. *Journal of documentation* 62: 350-71.

HCIL. 2003. *Treemap*. University of Maryland. Human-Computer Interaction Lab. Available: http://www.cs.umd.edu/hcil/treemap/

Hjørland, Birger. 2002. Domain analysis in information science: eleven approaches – traditional as innovative. *Journal of documentation* 58: 422-62.

Jacob, Elin K. 2004. Classification and categorization: a difference that makes a difference. *Library trends* 52n3: 515-40. Available http://findarticles.com/p/articles/mi_m1387/is_3_52/ai_n6080402/

Kosara,Robert and Miksch, Silvia. 2002. Visualization methods for data analysis and planning in medical applications. *International journal of medical informatics* 68n1-3: 141-53.

Kwasnik, Barbara H. 1999. The role of classification in knowledge representation and discovery. *Library trends* 48n1: 22-47. Available: http://findarticles.com/p/articles/mi_m1387/is_1_48/ai_57046525/

Lamping, John et al. 1994. Laying out visualizing large trees using a hyperbolic Space. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, 1994, pp. 13-14.

Mills, Jack. 2004. Faceted classification and logical division in information retrieval. *Library Trends* 52n3: 541-570. Available http://findarticles.com/p/articles/mi_m1387/is_3_52/ai_n6080403/.

Munzner, Tamar. 1998. Exploring large graphs in 3d hyperbolic space. *IEEE computer graphics and applications* 18n4: 18-23. Available http://graphics.stanford.edu/papers/h3cga/

Osińska, Veslava and Bala, Piotr. 2008. Classification visualization across mapping on a sphere. In: *New trends of multimedia and network information systems*. Amsterdam: IOS Press, pp. 95-107.

Osińska, Veslava and Bala, Piotr. 2009. Nonlinear approach in classification visualization and evaluation. In: *New perspectives for the dissemination and organization of knowledge: Proceedings of the IX Spain Group ISKO Congress 11-13 March Valencia, Spain*. pp. 222-31. Available http://dialnet.unirioja.es/servlet/fichero_articulo?codigo=2923178&orden=0

Pfeffer, Magnus et al. 2008. Visual analysis of classification systems and library collections source. In *Proceedings of the 12th European conference on Research and Advanced Technology for Digital Libraries Lecture Notes In Computer Science*, vol. 5173. Berlin; Heidelberg, Springer-Verlag, pp. 436-39.

Randelshofer, Werner. *Visualization of large tree structures*. Available http://www.randelshofer.ch/treeviz/index.html.

Scneiderman, Ben. 1998-2009. *Treemaps for space constrained visualization of hierarchies*. Last updated June 25th, 2009 by Catherine Plaisant. Available http://www.cs.umd.edu/hcil/treemap-history/

Shera, J. H. 1965. Classification as the basis of bibliographic organization. In *Libraries and the organization of knowledge*. Hamden, CT: Archon.

Taylor, Arlene G. 2006. *Introduction to cataloging and classification*. 8th ed. Englewood, Colorado: Libraries Unlimited.

All URLs are last checked in February 2010.