New Methods for Visualization and Improvement of Classification Schemes: The Case of Computer Science

Veslava Osińska* and Piotr Bala**

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

** Department of Parallel and Distributed Computing, Faculty of Mathematics and Computer Science, N. Copernicus University, Chopina 12/18, 87-100 Toruń, Poland, <bala@mat.uni.torun.pl>

Veslava Osińska received her Ph.D. in Library and Information Science from Nicolaus Copernicus University in Torun (Poland), where she lectures on Information and Communication Technology as well as Computer Graphics. Her current research activity involves the information visualization methods with particular consideration knowledge domain visualization (KDViz). She has applied her physics and computer science background to study nonlinear properties of information streaming. She is a member of ISKO Polish Chapter as well as Polish Computer Science Society.

Piotr Bala, Professor at Faculty of Mathematics and Computer Science N. Copernicus University, director of Department of Parallel and Distributed Processing. He obtained a PhD in Physics in 1993 in the Institute of Physics, N. Copernicus University and, in 2000, habilitation in physics for development and application of new methods to study enzymatic reactions in biological systems. He has published over 60 papers. His main research interest is in the development of parallel and grid tools for scientific applications. He has been involved in a number of national and European ICT projects.

Osińska, Veslava, and Bala, Piotr. **New Methods for Visualization and Improvement of Classification Schemes: The Case of Computer Science.** *Knowledge Organization, 37(3)*, 157-172. 48 references.

ABSTRACT: Generally, Computer Science (CS) classifications are inconsistent in taxonomy strategies. It is necessary to develop CS taxonomy research to combine its historical perspective, its current knowledge and its predicted future trends – including all breakthroughs in information and communication technology. In this paper we have analyzed the ACM Computing Classification System (CCS) by means of visualization maps. The important achievement of current work is an effective visualization of classified documents from the ACM Digital Library. From the technical point of view, the innovation lies in the parallel use of analysis units: (sub)classes and keywords as well as a spherical 3D information surface. We have compared both the thematic and semantic maps of classified documents and results presented in Table 1. Furthermore, the proposed new method is used for content-related evaluation of the original scheme. Summing up: we improved an original ACM classification in the Computer Science domain by means of visualization.

1.0 What is the problem with Computer Science (CS) taxonomy?

Computer Science (CS) taxonomy should cover every major aspect of computer science and technology as well as the latest technology with practical applications. Quite often, the evolution of the science disciplines overtakes the development of their classifications, especially for the newest subdisciplines and subdivisions. This situation is evident in the domain of Computer Science, which develops continuously. Computer Science emerges from computing needs. In an earlier phase, we found the development of numerical methods for purely theoretical and mathe-

matical tasks. When scientific data became more numerous and complex, information systems supported data processing and database management. With the introduction of PCs, the existing operating systems and graphic interfaces were multiplied. Gradually, CS has taken on an applied character and has been exploited, among others, by physics, astronomy, biology, geography and art. Today, the focus of CS goes beyond properties of electronic devices and also encompasses the process of human understanding and the design of user-directed components. CS embraces artificial intelligence, including problems in robotics, neuroinformatics, computer vision and so forth. For research into current trends in CS, we have to combine a historical perspective with current knowledge and with predicted future trends including possible breakthroughs in information and communication technology.

The changes in the field leave traces in domain taxonomy. CS classification schemes have been developed and used to categorize specialist literature. But these are not coherent and strictly logical relative to their development. A practical approach to describe the development of the field of CS is to take a closer look at CS classification schemes. Any weakness in these schemes might be cause for less effective information retrieval, in particular, if we compare the use of subject categories with a keywords/terms search. In this paper we address this problem and present a new way to visualize both classification schemes and keywords. In particular, we use the expert knowledge of the author of documents, which are codified in work and in the allocation of pre-defined classification schemes, to propose an alternative navigation through classification schemes.

There are different reasons for the inconsistent strategy of CS taxonomy, which we will discuss in the following. A first source is the old problem of computer scientists (Denning 2005), which is best expressed by the question: Is computer science "science?" The relative youth of the field (the first digital computers were invented in the mid-20th century) and many roots in other disciplines like mathematics, physics, electronics cause us to question the place of CS in the space between the natural sciences and engineering. Originally, the natural sciences included physics, chemistry, and biology; fields that addressed natural world objects. Computer science can be seen partly as a branch of mathematics, proving specific algorithms to solve problems. However, computational or simulation models of computer applications also describe complex natural phenomena and are therefore also part of the natural sciences. Often, CS models are used to simulate artificial worlds and complex phenomena, which escape direct experimentation (e.g., the climate or the weather). But computer scientists also create and experiment with information systems and develop methods and technologies to design, realize and operate them. This part of CS has been labelled information science. Moreover, it is actually difficult to pinpoint how much of CS is engineering and how much is theoretically-directed. CS partly penetrates other knowledge domains, and partly constitutes its own field with different specializations. This makes a strict definition of CS as a field and the identity of CS professionals problematic. Controversies about these hot topics can be found back in CS community forum debates (CFCS 2004, Constable 2000, CSAB 1997).

In computer science, there exists a limited terminological agreement. Culture and history determine the alternatives computing science (in the USA) and informatics (used more often in Europe: for example, we can mention the Informatics Europe association whose goal is to foster the development of quality research and teaching in information and computer sciences in Europe). The term most relative to informatics is cybernetics (used as an informatics equivalent in the former Soviet Union and some countries of East Europe), which seeks to develop general theories of communication within complex systems (Umpleby 2000).

A majority of users confuse computer science with the more accessible areas of computer maintenance, such as information technology (IT), or think that it relates to their own experience of computers, which typically involves activities such as gaming, web-browsing, and word-processing. For example, the question of the separation between CS and IT has been discussed recently and initiated at the high school educational level (Syslo and Kwiatkowska 2005).

The expansion of new computer technologies causes the development of branches with narrower or wider ranges of specialization. An example is artificial intelligence (AI) which has achieved its greatest success since the 90s. Knowledge from AI was adopted in the technology industry, providing the heavy lifting for logistics, data mining and medical diagnosis. Cognitive science creates a new connection between CS and human psychology, best visible in research on brain-computer interfaces. AIis an example for a field which started as a CS branch, but has evolved more and more into an enormous self-determining field in its own right (Kingston 2002).

This complex history and current development of a relative new scientific field which is so much "in-between" other fields also finds an echo in taxonomies and ontologies. We still see the influence of a variety of hand-made ontologies, which provide browsing structures for subject-based information in Web services specialized in CS and IT domains (Sosinska-Kalata 2002). On the other side, there also exist automated classification techniques (Golub 2006), but these can result in a proliferation of schemes.

In contrast to various dissimilar and specifically tailored ontologies, the most well-known subject classification scheme in the CS domain is the Computing Classification System (CCS). This system uses as main headings or classes: hardware, software, computer systems, information and data, mathematics of computing, theory of computation, methodologies, applications, and computing milieux (Osinska 2005). The CCS was created by the oldest informatics society, the Association for Computing Machinery (ACM), as a response to a growing collection of science publications concerning the CS domain. The first version was published in 1964. The system is still in use and has, of course, been periodically updated. In particular in the last two years, both the CCS structure and interface have been changed significantly. The current CCS scheme is based on the original taxonomy of computer science and engineering, which was published by the American Federation of Information (CCS Report 1998).

The full current classification tree can be studied online (The ACM Computing Classification System 1998). Its distinct structure, as well as the ACM Digital Library accessible online resources, facilitate an analysis of this classification scheme. Their linear organization depicts the intersection between subclasses; this means the classification does not meet the exclusiveness principle. The traditional way to visualize classifications is with Dendrograms. Their main disadvantage is a limited identification of overlapping crossing classes.

In this paper, we propose a specific technique to reveal the hidden links between classes while keeping their correlations according to the Dendrogram structure. We decided to map the CCS classification tree onto a nonlinear space–-a spherical surface. In particular, we applied a visualization method the details of which are described elsewhere (Osinska and Bala 2008). In this paper, we concentrate on the discussion of how such a new visualization layout can be used for an evaluation and subsequent modernization of the current CCS classification scheme. Before we introduce our own method, we will describe in the next section the state of the art in knowledge visualization. We review the main developments in this field that form a kind of methodological background for our own approach. In section 3, we then introduce our method in more detail, using the example of CS taxonomies; in section 4, we attach the given visualization maps and interpret them. We also compare the structural organization of thematic categories from graphical layouts and the original classification scheme. In the following section, all stages of the visualization process including results and future research plans are summed up.

2.0 How to visualize a scientific field?

For the last decade, the visualization of scientific domains (not to be confused with scientific visualization focused on measured data) has been expanded methodologically and applied to a wide spectrum of different disciplines and fields. Different research projects join the effort to create a cartography of knowledge based on modern computational algorithms and cognitive insights. Concepts and methods from scientometrics on how to measure and analyse the sciences going back many decades, have been taken up by computer scientists interested in the visualization of scientific knowledge (Börner et al. 2003; Bollen et al. 2009; Boyack et al. 2002; Holloway et al. 2005 Chen 2006). One of the most active researchers for the last decade in information visualization (Infoviz), Chaomei Chen enumerates modern visualization techniques and methods in his book (2006). He first introduces the notion for a newly emergent field "knowledge domain visualization (KDViz)" which contributes to a better understanding of the structure and dynamics of knowledge. Another, and even older, label for the mapping of science (not widely adopted though), is "scientography," proposed by Eugene Garfield (1994). A detailed historical review of the ways that the sciences have been visualized as a system can be found in Moya-Anegón et al. (2004). The early work on science mapping used small sets of databases (conference proceedings, grant data, communities papers). On a larger scale, representations of different scientific domains were performed using bibliographic data from ISI (Institute for Scientific Information, actually now known as Thomson ISI) and databases such as the *Journal Citation Reports* (*JCR*), the *Science Citation Index* (*SCI*), the *Social Sciences Citation Index* (*SSCI*), and the *Arts and Humanities Citation Index* (*AHCI*). Eugene Garfield, the founder of the ISI,

explains in his essays (1998) how we can generate global science maps and how we can discover both research fronts and the interests of researchers. Using a series of chronologically-sequenced maps, it is possible to track the evolution of scientific domains—as defined by Eugene Garfield using the notion of longitudinal mapping (1994).

Visualization methodologies provide a spatial representation of interrelationships between investigated units. Scientific publications, journals, authors, or subject categories can be used as units of analysis. Citation analysis proposed independently by Irina Marshakova (1973) and Henry Small (1973) became the most popular method used for identifying thematic trends and the predominant research areas in a given field. Nowadays, knowledge domain mapping goals can be extended to information retrieval, to science trends monitoring or to science and technology management politics (Boyack et al. 2002). Birger Hjørland (2002, Hjørland and Albrechtsen 1995) has pointed to the wider social aspects of domain analysis as a new approach of information science.

While more and more scientific publications can be accessed on the Web, KDViz researchers look for new, more dynamic measures and approaches to domain visualization. Besides bibliographic similarities, they can work out linguistic, text mining algorithms, in order to find similarities between documents be it topics- and/or semantics-based. Moya-Anegón and associates (2004) obtained an effective way of science category mapping by using co-cited ISI category assignments. Katy Börner—the author of numerous publications about KDViz with her InfoVis colleagues, has analysed and visualized the network of articles in English Wikipedia (Holloway et al. 2005). They employed a link analysis technique and generated a base map using a measure of similarity of categories. Johann Bollen's group (2009) has exploited the advantages of log datasets over citation data by measuring clickstreams: sequences of user requests. Scholarly web portals record detailed users' logs at a scale exceeding the number of citations sets. Furthermore, log datasets reflect the activities of a larger community as well as record the interactions between users of scholarly portals.

Contemporary information which originates from the Web is evermore complex and multidimensional. The effectiveness of visualization methods for information retrieval and navigation depends among others on how the dimensions of the usual multidimensional space of knowledge objects are reduced, and on which rendering techniques are applied to represent the data on a screen in the most legible way. Webometrics and contemporary scientometrics methodologies need to be connected to other interdisciplinary and fast-moving knowledge domains. Examples are KDViz techniques, which evolved from statistical algorithms like clustering, multidimensional scaling (MDS) and factor analysis, but also self-organizing maps (SOM), models of graphs (Pathfinder Networks), and complex networks (Leydesdorff 1987; Smiraglia 2008; McCain 1998; Börner et al. 2009; Scharnhorst 2003).

Despite the ISI citation indexes still constituting the preferred scientific bibliographic data sets, they do not adequately cover the journal literature in all subfields (Moed and Visser 2007). Friedemann (2008) presented on the European Computer Science Summit (ECCS) discusses why ISI data are "harmful" in bibliometric evaluations of the Computer Science domain. Un the ISI databases, CS journals, in contrary to natural sciences, constitute only a small piece (about 4 percent) of all records. The ISI data does not include most CS conference and workshops proceedings where a majority of significant scientific papers appear. Another problem concerns the excessive disproportion of CS research in the USA and EU in the ISI databases. Therefore, Friedemann (2008) has suggested that a mapping of the CS domain should combine traditional databases such as ISI, NLCS, ACM, IEEE, as well as Scopus, Google Scholar and so forth. These arguments and the specifics of Computer Science as a field (as discussed in previous chapter) might explain why this discipline has not been visualized so far. Obviously, we need to study not only quantitative properties of the CS domain but relate these findings also to a qualitative interpretation.

Our research includes an analysis of Computer Science literature data on every level of classification hierarchy. We start by analyzing and mapping objects on the document level. We extract information from these documents and receive a topology of a classification on a high agglomeration level, namely of classes and subclasses. This classification topology (as a collective effect) then determines the locations of single documents in it. For the actual visualization algorithms and the optimal representation of a given classification on a 3D space, we used the law from molecular physics (Osinska and Bala 2008).

We are aware of the role of contemporary design rules and user requirements for an intuitively understandable but also aesthetically pleasing presentation. An interesting example for the meeting of art and science maps can be found in the attractive picture

"Hypothetical Model of the Evolution and Structure of Science" by Daniel Zeller (Zeller 2007). This art object suggests and seduces us to perform a graphical processing of visualization results which clearly puts the emphasis on the emergence of new fields and their deep roots back to history. The CS domain structure map we present in this article shares some topological features with this art representation. However, it is the combination of co-occurrence of classifications (as externally defined features) and keywords (as internally defined) which make our method original compared with others. In this way, social patterns not only of scientific activity, but also of the efforts of editors can be discovered and verified. Finally, we use the visualization results as an evaluation interface of the original classification scheme.

3.0 Visualization process: new insights into CS

3.1 Methodology of data mapping

In this paper we use a new graphical representation of an original classification scheme. The details of this method have been described elsewhere (Osinska and Bala 2008). We repeat in this article only the main elements of the method referring to our concrete example. We started with testing data sets retrieved from the digital library, which have been classified according CCS. The classification tree is characterized by three levels of hierarchy: the main classes and two levels of subclasses. Please note that the highest population of data can be found on the lowest levels. In other words, the majority of documents is classified in a very fine grained way. We did not notice features such as adequacy and disjointedness of subclasses. If some sublevel nodes split semantically, the documents appear in both nodes. This feature was employed in our current methodology. In other words, a document can carry different classifications from different parts of the classification tree; its location on the tree is not unique, and a document can appear multiple times. The appearance of different classes in the classification of a document allows us to define links between classes. We call these document attributes co-classes. We assume that the topic similarity between co-classes is proportional to the number of recurrent documents. The closer semantically two subclasses are, the more they include common articles. These pairs of classes lead to cross links in the typical Dendrogram tree. Inversely dissimilar subclasses contain no common data.

By counting and normalizing the number of common documents for every pair of classes and subclasses, we can construct a similarity matrix. In our database we found 353 classes and subclasses. The dimension of the square matrix was equal to the number of all occurrences in the data set's classes and subclasses, i.e. 353×353. In order to decrease this dimension we used an MDS 3D plot. As a target space, we selected the sphere surface. Spherical surfaces have a few special properties which make them particularly suitable as interface. The sphere, a symmetrical figure, is ergonomic for both browsing and navigational processes. The curved surface has no edges and offers less distortion than a rectangular plane does in the distribution of classes nodes.

3.2 Dataset

The last version of the CCS System was published by ACM in 1998 and is still actualised on the fly. The digital ACM library includes a significant collection of abstracts and full-text scientific publications (1.4 million. text pages), ACM journals and conferences proceedings. The classification tree is restricted to three letter-and-number-coded levels in order for it to reflect accurately the essential structure of the discipline over an extended period. An uncoded fourth level of the tree, subject descriptors, provides sufficient detail to cope with new developments in the field. 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 Milieus

Every publication, besides this main classification, may be described with additional classes. Detailed instructions from CCS editors give authors the information to clearly classify their documents (How to classify works …). Authors have to describe the document's categories, keywords and implicit subject descriptors. Browsing document abstracts, the user can see the automatically generated tree of main and additional classifications. Figure 1 illustrates sche-

matically how one can access a document's metadata. In the example, three co-classes with symbols: I.4.8, F.2.1 and I.5.2 are ascribed to an article titled "Detection of planar motion objects." The first symbol indicates the main class, and the next two indicate additional classes. The presence of multiple subclasses signifies a wide semantic context. Hence authors, together with editors, contribute to the semantic topology of the documents set.

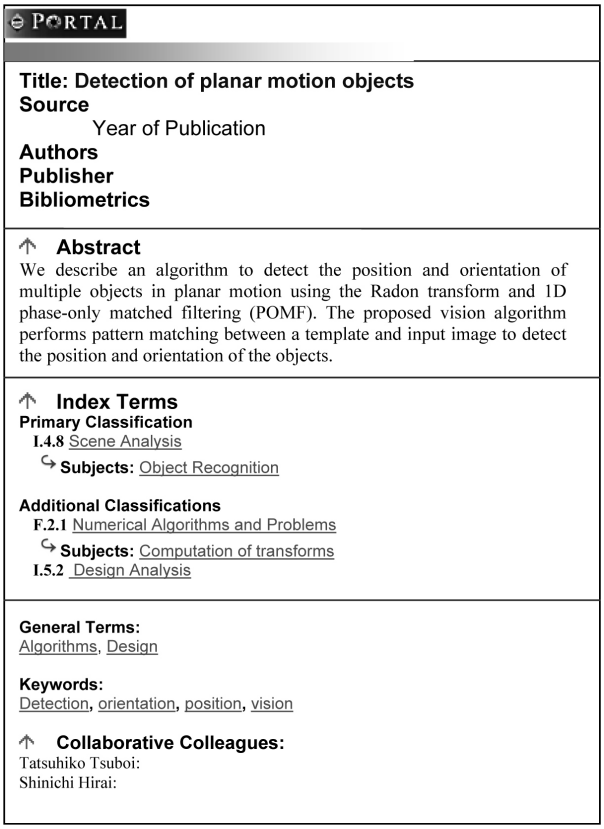

Figure 1. Schematic presentation of document metadata in ACM Digital Library. As themes categories analysis units: classes symbols and keywords were used.

3.3 Architecture of experiment

A database of documents' metadata was collected by a PHP5 application; a scanner running on an Apache Web server. Metadata were crawled on Websites' content in the following order:

- primary and additional classification symbols
- keywords
- global terms
- data of publication
- URL address.

Statistics of the data were generated with support of VBA (*Visual Basic for Application*). A similarity matrix has 353 dimensions and consists of co-occurrence numbers of classes. The matrix was then normalized to the documents' number of classes. The VBA engine was then used to apply an MDS algorithm in order to receive coordinates in 3D space. Further data processing and final visualization occurs in the Matlab environment. These procedures, steps, and tools are illustrated in Figure 2.

3.4 Data map processing

The graphical 3D representation of classification is shown in Figure 3a. The (sub)class nodes emerge from a dense pattern of coloured document nodes. In the initial stage of class visualization, we use three attributes: colour, luminosity and size of nodes. The colour indicates one of 11 main classes; the luminosity indicates the level of the tree; and the size indicates (sub)class population. From these class node coordinates we determine all the documents' positions (37343) on a spherical surface using topological centre rules. We assumed that the weights of primary and additional classification are 0.6:0.4. The reasons for this supposition are explained in Osinska and Bala's work (2009). The document nodes inherit the colour of the main class. We obtain 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. For the next step, we decide to conduct further research by using cartographic projections of the sphere surface (Figure 3b) because of the necessity of graphic processing and non-linear evaluation (Osinska and Bala 2009). In any case, the output image, can be employed as sphere texture.

4.0 Data analysis

4.1 Theme categories map

The data points form colour patches with different densities and sizes (see Figure 3b). Obtained clusters are characterized by a dissolved border, and steps to adjust the map were therefore required. The median filter is a non-linear technique, often used to remove noise from images. We have classified noise, for our purposes, as single, distant points which disturb the final pattern of clusters. The later verification of documents belonging to these separated spots confirms their low information value. The median filter and the next contour filter were applied for edge de-

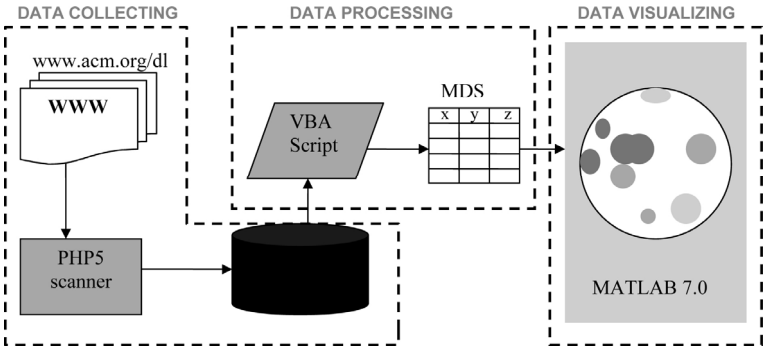

Figure 2. Overview of experiment's architecture.

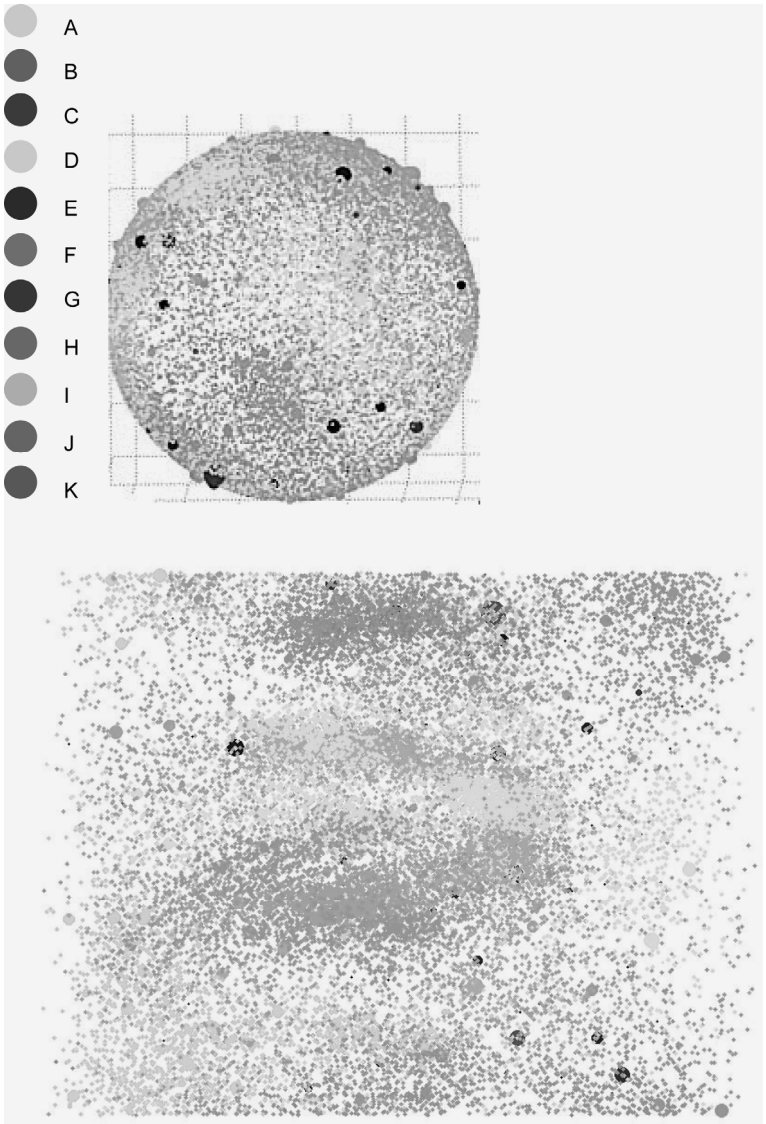

Figure 3. Classification mapping **a)** on a sphere surface. We can see 353 (sub)classes (in down layer) and documents nodes (upper layer); **b)** cartographic projection of sphere surface. Three attributes as: colour, luminosity and size of nodes were used. Color of ponts and glyfs identifies each of 11 main classes (legend).

tection. The final map of document clusters is shown in Figure 4. The colors of clusters relate to the 11 main classes of CCS. Now we can see not only the clear edges of clusters, but also such properties as overlap, and the splitting of clusters in mixed-colors areas which determine similar themes of the original classification tree. Ontologically-different hardware (B class) and software (D class) are distributed in opposite corners (or poles for the sphere). However, class C, the network category, is placed between them because both problems are represented.

4.2 Keywords map

Document topic identification within clusters by means of keywords was the second phase of analysis. The keywords within obtained clusters were used as the next units of analysis. We created statistical rankings of keyword frequency for each cluster. During analysis, it is important to consider any keyword within its neighborhood belonging to the same cluster. Clusters were captioned using the color of the proper main class, as one can see in Figure 5. Two variables, classes and keywords, were used separately in the mapping process. Thus, the keywords map now serves as a verification ground for the thematic one. The keywords map is organized logically according to linguistic content. Terms with similar meaning are located close to one another. In this context, we must discuss the local accuracy because it is hard to verify such a "visual thesaurus" on a global scale.

4.3. Thematic-semantic comparison

The essential feature of this layout relies on the primary classification of investigated data. While the CCS tree is characterized by three levels of details and unnumbered subject descriptors, a new organization

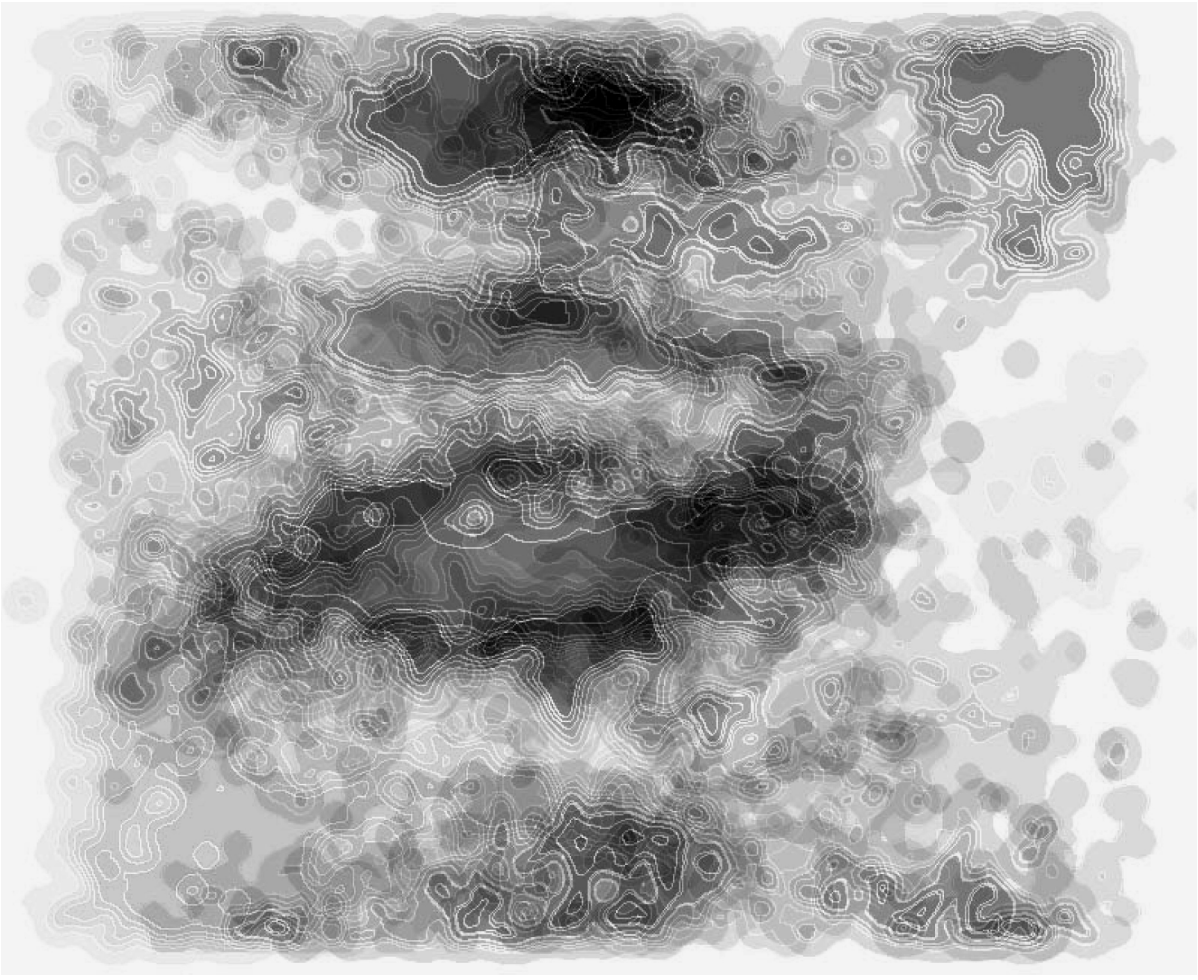

Figure 4. Contour map of CCS classification after image processing. Color of cluster identifies each of 11 main classes (legend).

Figure 5. Keywords map of CCS documents. Colour of font identifies each of 11 main classes (legend).

of articles is presented in simple non-hierarchical clusters. Information about the classification's hierarchy degrees, which carried a subtle structure has been lost. For all mapping processes, we use such metadata as primary and additional classification symbols and keywords. At the clusterization stage, the data about nested subclasses is effaced. The given visualization scheme was locked within one level hierarchy to reflect semantic proximity. When no overlap occurs between clusters we talk about the disjointedness of the "ideal" arrangement structure (Dal Porto and Marchitelli 2006). In the case of such model clusterization of a whole digital library collection, it may be safely concluded that computer science classification does not require as deep a hierarchy as CCS. Although most clusters are separated with a good resolution, some map areas are shared by them. Hence we need an additional approach for visualization evaluation.

Equivalent clusters of a new structure have the following attributes: population, size and data density. During semantic processing, we involved the next important meta-characteristic: keywords description of each cluster. We noted that the most cohesive results are obtained for high density clusters. The remaining scattered data objects are considered as information noise. It is motivating to investigate how the clusters on the semantic map described by keywords are covered with the primary thematic categories. This method constitutes an evaluation of subject categories organization in CCS classification.

We compared these two schemes in regards to thematic-semantic consistency. Table 1 represents some characteristic examples of good matching. The subclasses' symbols and topics in the second and third columns are set against keywords in the last column. Because the level of details in the first structure is not sufficient for the identification of articles, subject descriptors have been added. Clusters are sorted by the number of documents with described keywords. Only part of documents within each cluster are represented in the keywords sequence analysis because not all authors include these metadata in their publications.

Main Class	Cluster	CCS subclasses		Subject descriptors		Rank of Keywords
C	1 up 1128 / 1717	C	Computer Systems Organization	Hardware/software interfaces Instruction set design Modelling of computer architecture System architectures Systems specification methodology Adaptable architectures **Cellular architecture (e.g., mobile)** Data-flow architectures **Frame relay networks** Network topology **Wireless communication**	RISC/CISC, VLIW architectures Array and vector processors Associative processors Connection machines Interconnection architectures (e.g., common bus, multiport memory, crossbar switch) Heterogeneous (hybrid) systems **Data communications** Open Systems Interconnection reference model (OSI) **Security and protection (e.g., firewalls)** Circuit-switching networks ISDN (Integrated Services Digital Network) **Network communications** Packet-switching networks Store and forward networks	**Wireless LAN**
		C.1	Processor architectures			**Sensor networks**
		C.1.1	Single Data Stream Architectures			Quality of service
		C.1.2	Computer Systems Organization			**802.11**
		C.1.3	Processor architectures			Performance
		C.2	Computer-communication networks			**ad hoc networks**
		C.2.1	Network Architecture and Design			**security**
						mobile ad hoc networks (MANETs)
						Internet
						energy efficiency
						mobility
						routing
						multicast
						Fault tolerance
						Intrusion Detection
						Broadcast
						Scheduling
Computer Systems Organizations	1 down 371 / 567	C	Computer Systems Organization	**Network management** Network monitoring **Public networks** **Distributed applications** **Distributed databases** Network operating systems	Process control systems Real-time and embedded systems Smartcards Design studies **Fault tolerance** Modelling techniques **Performance attributes** **Reliability, availability, and serviceability**	**peer-to-peer**
		C.1	Processor architectures			**distributed computing**
		C.1.1	Single Data Stream Architectures			**grid computing**
		C.1.2	Computer Systems Organization			**quality of service**
		C.1.3	Processor architectures			Scheduling
		C.2	Computer-communication networks			**fault-tolerance**
		C.2.1	Network Architecture and Design			
		C.2.3	Network Operations			
		C.2.4	Distributed Systems			
		C.2.m	Miscellaneous			
		C.3	Special-purpose and app-based system			
		C.4	Performance of systems			
		C.5	Computer system implementation			
		C.m	Miscellaneous			
C	2 318 / 452	C.2.2	Network Protocols	**Protocol architecture** **Protocol verification** **Routing protocols**	Applications (SMTP, FTP, etc.) Open Systems Interconnection reference model (OSI)	**Routing**
		C.5	Computer system implementation			wireless networks
						Ad hoc networks
						Quality of Service
						Performance evaluation
						Sensor networks
						TCP
						mobile ad hoc networks
						Internet
						Security
						protocols

Main Class	Cluster	CCS subclasses		Subject descriptors		Rank of Keywords
C	3 214 / 345	C.1.2	Computer Systems Organization	Client/server Process control systems **Real-time and embedded systems** **Signal processing systems** **Smartcards**	Open Systems Interconnection reference model (OSI) Microprocessor/microcomputer applications Array and vector processors	**Embedded systems**
		C.1.m	Miscellaneous			**FPGA**
		C.2.0	Computer-communication networks			Performance analysis
		C.2.4	Distributed Systems			sensor networks
		C.3	Special-purpose and app-based system			**smart card**
						wireless
H	1 1917 / 2747	H.1	Models and principles	Software psychology Access methods **Statistical databases** **Clustering** **Query formulation** Relevance feedback **Retrieval models** **Search process** Recovery and restart Dissemination Systems issues **User issues** Desktop publishing Spreadsheets Word Processing Natural language Prototyping Training, help, and documentation **User-centered design** **Image databases**	**Data mining** **Human factors** **Human information processing** **Information filtering** **Scientific databases** Interaction styles (e.g., commands, menus, forms, direct manipulation) Graphical user interfaces (GUI) **Information browsers** Information networks **Performance evaluation** User profiles and alert services **Web-based services** **Animations, Video** **Audio input/output** **User interface management systems** Artificial, augmented, and virtual realities **Evaluation/methodology**	**Database**
		H.1.2	User/machine systems			Visualization
		H.2.2	Physical design			**Data mining**
		H.2.8	Database applications			**Information retrieval/seeking**
		H.3	Information storage and retrieval			**Web Services**
		H.3.3	Information Search and Retrieval			**query expansion/processing**
		H.3.4	Systems and Software			Ontology
		H.3.5	Online information services			context information/awareness
		H.3.7	Digital libraries			**Clustering**
		H.4	Information systems applications			decision support/making
		H.4.1	Office Automation			**search strategy/engine**
		H.4.3	Communications Applications			semantic web
		H.5.1	Multimedia Information Systems			**HCI/human-robot-interaction**
		H.5.2	User Interfaces			**Multimedia**
		H.5.4	Hypertext/Hypermedia			**Evaluation**
		H.5.5	Sound and Music Computing			**user study**
						user interfaces
						Web search
						Knowledge Management
Information Systems	2 449/597	H.1	Models and principles	General systems theory Information theory Value of information Abstracting methods Dictionaries **Indexing methods** Linguistic processing Thesauruses	**Decision support (e.g., MIS)** Logistics **Data description languages (DDL)** **Data manipulation languages (DML)** **Database programming languages** **Query languages**	**Database**
		H.1.1	Systems and Information Theory			**query processing**
		H.2	Database management			XML
		H.2.1	Logical design			Ontology
		H.2.3	Languages			**Decision support**
		H.2.4	Systems			**Indexing**
		H.2.5	Heterogeneous databases			Information retrieval
		H.2.m	Miscellaneous			semantic web
		H.3	Information storage and retrieval			Performance analysis
		H.3.1	Content Analysis and Indexing			
		H.4.2	Types of Systems			
		H.4.m	Miscellaneous			
	3 303/471	H.1.1	Systems and Information Theory	General systems theory Information theory Value of information **Data models** Ergonomics Natural language Theory and methods **Schema and subschema** **Image databases** **Scientific databases**	Abstracting methods Dictionaries Indexing methods Linguistic processing Thesauruses **Decision support (e.g., MIS)** Evaluation/methodology Graphical user interfaces (GUI) Interaction styles (e.g., commands, menus, forms, direct ma-	**Decision support system**
		H.2	Database management			**Databases**
		H.2.1	Logical design			**XML**
		H.2.5	Heterogeneous databases			Knowledge management
		H.2.8	Database applications			Visualization
		H.3	Information storage and retrieval			Fuzzy sets
		H.3.1	Content Analysis and			Ontology

Main Class	Cluster	CCS subclasses		Subject descriptors		Rank of Keywords
			Indexing	**Spatial databases and GIS** **Statistical databases**	nipulation) Screen design (e.g., text, graphics, color)	
		H.4.2	Types of Systems			Multicriteria decision
		H.5	Information interfaces and presentation			data warehouse
		H.5.2	User interfaces			Conceptual modelling
						Data models
	4 163/222	H	Information Systems	Evaluation/methodology Organizational design Asynchronous interaction Synchronous interaction **Collaborative computing** Theory and models	**Computer-supported cooperative work** **Web-based interaction**	**Collaboration/collaborative learning**
		H.5.3	Group and Organization Interfaces			**Social computing**
						Wikipedia
						Knowledge management
						Communication
						computer-mediated communication
						Awareness
						Visualization

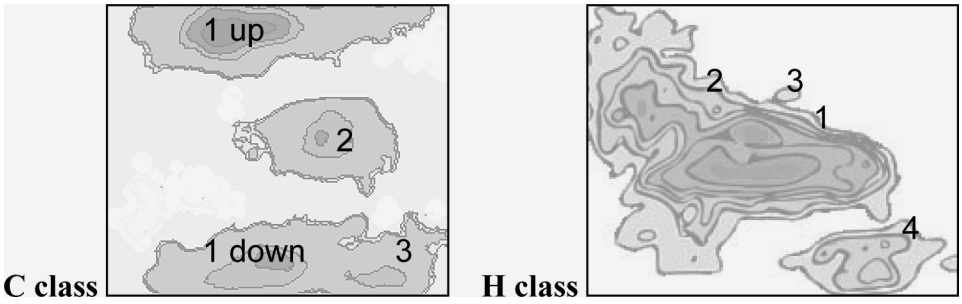

Table 1. The comparison of topic organization within classification CCS and experimental clusterization.

As we can see in the schemes attached to the tables, three clusters are formed from the nodes of documents belonging to the main class **C** (computer systems organizations). Two investigated units of attributes, such as the subject's descriptors and keywords, are independent. The similar data in the last two columns is marked by bold text. The cluster "1 upper" (scheme under the table) deals with computer communication network problems, especially wireless networking with frequent citing of its decentralized type—"ad hoc networks." A significant part of the data is derived from articles about mobile technology. The common category phrase 802.11 represents a set of standards for implementing a wireless local area network. The keyword "broadcast" can be generalized as a data communication mode. The authors also bring up "*security networking*" so this keyword is present in both datasets. It should be noted that the subject descriptors do not list current networking field terms, for example: "LAN, routing, ad hoc, Ethernet, broadcast." Instead, the obsolete technology ISDN is used. The cluster "1 down" additionally (a new symbols of subclasses were arrived) refers to "distributed computing," and its modern form—"grids." No exact subject descriptor is provided but close categories such as "distributed application, distributed database" and "network management" can characterize these topics. Additionally, the "quality of service" one can relate to "reliability" and "serviceability." Cluster *2* specializes in a "network protocols" and "routing," cluster "3 – embedded systems."

The next tested class **H** (information systems) includes the wide spectrum of information science topics. It is not easy to define the main field of the biggest cluster "1." Here there are articles concerning data mining, information retrieval, clustering, Web services, and most types of databases. Many terms deal with the study of interaction between users and computers: "Human- Computer Interface, Human Factors, User Study." But no subject equivalent for keywords:"ontology" and "knowledge management" is found. Accordingly, these words do not occur as major categories in keyword sequences. Cluster *2* can be regarded as an outcome of query processing or decision-making publications. The database languages such as *SQL, DDL* as well as "indexing" are included

in the "database" theme. Following in the table, cluster "3" adopts the features of the first cluster because of its close localization. A new issue: "schema and subschema" can be related to the keyword "XML." The category of subclass "H.2.5. Heterogeneous databases" is no longer used and it is labelled as "revised" in the CCS tree. Generally, cluster "3" is noticeably described by "knowledge management, ontology," and "semantic web" topics. But these categories are registered within class **I.** (methodologies) subclass "I.2. artificial intelligence." As the visualization map shows, these two clusters are close to one another thematically. This observation proves that CCS does not fully correspond to the current stage of Computer Science development. Moreover, there is no subject descriptor "Visualization" or even a related term in classification. Many articles with the major topic "visualization" can not be classified precisely. The thematic direction of cluster "4" we characterized as "social computing: Wikipedia, collaboration," and "collaborative learning."

By comparing subject descriptors and keywords of clusters, we are able to abstract salient features of their thematic organization and thereby name them suitably and arrange them within consistent main classes.

5.0 Visualization results summary

Summarizing the experiment's results for convenience, we can specify four main phases of data analysis: the visualization of results on the sphere, the mapping with image processing, the keywords clustering, and the classification modernization.

5.1 Visualization on the sphere surface

The crucial achievement of this current work is a logical visualization of documents from the ACM Digital Library. We based our similarity metrics on the number of co-classes. We used such metadata as primary and additional classification symbols. Because of dataset magnitude, we investigated only the collection of articles published in the year 2007. The final number of classes and subclasses was 353, and the dataset consisted of 37,543 documents. A spherical surface was chosen as our preferred mapping and navigation space. By foregoing linear methods, it was possible to represent data graphically and to keep similarity. To reduce data matrix dimensions, we used an MDS technique. Uniform distribution of document nodes on our spherical mapping surface proves that this is a proper strategy for the visualization of classification trees and digital library collections.

5.2. Visualization maps

Sphere rotation and zoom provide easy browsing of data and observation of their relations in a hyperbolic space according to the principle "focus+context" (the technique "focus+context" implemented in the interfaces provides the user both with an overview (context) and with detailed information (focus) simultaneously). In order to perform further analysis, we used the projection of this data onto a plane. In this particular case we have used a cartographic projection to flatten the spherical surface. The obtained map shows a different concentration of data points around class nodes. As a digital image with a highly complex structure, the map requires use of nonlinear processing. Nonlinear graphic filtering techniques were applied to the maps. To remove noise and detect clusters edges, we applied median and contour filters sequentially. The algorithms used gave crucial information about the main classes' frontiers.

5.3 Semantic map

In this stage we used the next attribute of documents—keywords. The rankings of keywords for all data points in clusters were calculated. Clusters were designated by the most frequently-used keywords sequences. A semantic map (shown in Figure 5) of keywords obtained this way reveals important properties such as local accuracy. The semantic map was then used in the next process—evaluation of the existing classification scheme. Accuracy at this point means similarity in both paradigmatic and intuitive comprehension of themes. It should be taken into consideration that the keywords are an effect of author's competence and exactness—however, this human factor can introduce fault.

5.4 A "new" classification.

Next, our work was oriented towards our keyword lists' confrontation with the existing classification scheme. The resulting Table 1 presents the comparison of topic organization within CCS classification and its given clusterization. We randomly chose to analyze two main classes: **C.** (computer systems organizations) and **H.** (information systems). For precise identification of article topics, the subject-descriptor sequences within any subclass were ana-

lyzed. By comparing the subject descriptors and keywords of any cluster, it is possible to ascertain rules of its thematic organization, including their common features. Keyword frequency can provide information about the cluster's association degree with any given topic.

The subsequent study of Table 1 concerns some noticeable topic gaps within CCS. For example no subject descriptor "Visualization" or relative was found in the classification scheme, although a lot of articles utilize this as a major topic. It is possible to discover obsolete thematic categories such as "Heterogeneous databases, analog computers" etc. Especially noticeable is the non-practical organization of the class "I. Methodologies." Namely, the "knowledge organization" theme appears only within subclass "I.2. artificial intelligence." J. Kingston proposed an extension of AI subjects to the ACM classification scheme by means of multi-perspective analysis (Kingston 2002). All these details reveal that CCS requires a new systematic approach in order to achieve a close correspondence to the current stage of computer science development. Particular corrections of this scheme are made by editors locally through, for example, using the labels "new," "revised," or "no longer used."

6.0 Conclusion

The obtained visualization maps depict the organization of the contemporary CS domain as it is reflected by scientific output. A majority of KDViz works use citation data. We proposed a new approach for visualization based on classification and by independently analysing thematic categories and keywords. We concentrated on the content properties of documents. In our case, classes, not citations, carry the information about theme categories. As explained above, one motivation for our approach was some lack of coverage of research in CS in traditional databases such as the *Web of Science*. We decided that for our goal to study the internal structure of computer science, the journal citation mapping approach was not suitable. CCS editors decide on articles' categorization ultimately, and keywords sets are defined by authors solely. The first can be seen as a kind of specialist annotation which is located on a more institutional, already aggregated level of scientific activity, and which is to a certain degree separated from the original research practice. The latter is codified by the authors using keywords, and making use of the classification structure offered a-priori. The innovative approach of our methods lies in a combination of both assignment activities. In other words, we get access to the friction between actual (individual or group-based) research practices and agreed epistemic categories on the level of the whole scientific community. In this tension we might see indications for change in the scientific field under observation. The main goals of the current research (besides achieving effective visualization) were the evaluation of the original classification scheme and its possible improvement.

Supposing that the outcome clusterization reflects the logical categorization of modern computer science literature, then, the covering of thematic-semantic categories within the clusters on the visualization map can inform us about the quality of organization of the input classification. Table 1 shows the comparison of topic organization in both schemas. As a result we see that the clusters which are grouped together in our visualization by means of keywords are effective. The main feature of the transformation of a classification space to a semantic space is a reduction of hierarchy levels. As shown here, the simplest hierarchy in the classification scheme is sufficient to create a rational classification scheme of digital library resources like ACM while preserving thematic similarities. This method could be applied in automatic classification tasks.

The model of clusterization which is obtained by the presented experimental procedure is independent from the initial CCS classification scheme. The visualization map which defines the organization of clusters—contrary to Dendrograms—is not linear. Thus we are entitled to describe it as a new or reorganized classification. With our method, we provide computer scientists with a practical tool to gain insight into CS research fronts and multidisciplinary trends for their domain. On the other hand, we still lack a qualitative assessment of the given visualization and its derived clusterization. Future research could focus on methods of similar article retrieval within the map according to a model of Birger Hjørland (2008). At present, preliminary tests have generated promising results.

In the future, we plan to repeat this research in cyclical periods every 10 years. By creating a longitudinal map of CCS classification, it will be possible to build a dynamic knowledge space for the Computer Sciences. Appropriate applications with animated layouts may demonstrate domain history and be able to predict which subfield will become far-reaching and which will decay. Eventually, we want to highlight that innovative methods can also be applied to other knowledge domains, especially multidisciplinary fields, and other databases.

References

The ACM computing classification system [1998 Version]. Valid through 2010. Available online at URL: http://www.acm.org/about/class/ccs98-html

Adkisson, Heidi P. 2005. Use of faceted classification. *Web design practices*. Available online at URL: www.webdesignpractices.com/navigation/facets.html.

Bollen, Johann et al. March 2009. Clickstream Data Yields High-Resolution Maps of Science. *Public Library of Science (PLoS ONE)* 4n3. Available online at URL: http://www.plosone.org/article/ info%3Adoi%2F10.1371%2Fjournal.pone.0004803.

Bonitz, Manfred. 1990. Information—knowledge—informatics. *International forum on information and documentation* 15n2: 3-7.

Börner, Katy and Scharnhorst, Andrea. 2009. Visual conceptualizations and models of science. *Journal of informetrics* 3n3. Available online at URL: http://arxiv.org/ftp/arxiv/papers/0903/0903.3562.pdf.

Börner, Katy et al. 2003. Visualizing knowledge domains. In Cronin, Blaise, ed., *Annual review of information science & technology*, v. 43. Medford, NJ: Information Today, pp. 179-255.

Börner, Katy et al. 2007. Network science. In Cronin, Blaise, ed., *Annual review of information science & technology*, v. 41. Medford, NJ: Information Today, pp. 537-607.

Boyack, Kevin W. et al. 2002. Domain visualization using VxInsight for science and technology management. *Journal of the American Society for Information Science and Technology* 53: 764-74.

Chen, Chaomei et al. 2001. Visualizing latent domain knowledge. *IEEE transactions on systems, man, and cybernetics.* Part C 31: 518-29.

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

Committee on the Fundamentals of Computer Science: Challenges and Opportunities, National Research Council. 2004. *Computer science: reflections on the field, reflections from the field*. National Academies Press. Available online at URL: http://www.nap.edu/catalog.php?record_id=11106#toc.

Computing classification system 1998: current status and future. 1998. Report of the CCS Update Committee. Available online at URL: http://www.acm.org/about/ class/ccsup.pdf.

Computing Sciences Accreditation Board. 1997. *Computer science as a profession*. Available online at URL: http://www.csab.org/comp_sci_profession.html. Retrieved on 2008-09-01.

Constable, Robert.L. 2000. *Computer science: achievements and challenges circa 2000*. Ithaca, NY: Cornell Univ. Press. Available online at URL: http://www.cs.cornell.edu/cis-dean/bgu.pdf.

Coulter, Neal et al. 1998. Computing classification system 1998: current status and future maintenance report of the CCS Update Committee. *Computing reviews*. New York ACM. Available online at URL: http://www.acm.org/about/ class/ccsup.pdf.

Dal Porto, Susanna and Martchitelli, Andrew. 2006. The functionality and flexibility of traditional classification schemes applied to a content management system (CMS): dacets, *DDC, JITA*. *Knowledge organization* 33: 35-44.

Denning, Peter. 2005. Is computer science science? *Communication of the ACM* 48, 4: 27–31.

Friedemann, Mattern. 2008. *Bibliometric evaluation of computer science –problems and pitfalls.* Available online at URL: www.informatics-europe.org/ECSS08/papers/mattern.pdf.

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

Garfield, Eugene. 1998-. *Essays/papers on "mapping the world of science."* Available online at URL: http://garfield.library.upenn.edu/mapping/mapping.html.

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

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

Hjørland, Birger. 2008. What is knowledge organization (KO)? *Knowledge organization* 35: 86-101.

Hjørland, Birger and Albrechtsen, Hanne. 1995. Toward a new horizon in information: domain-analysis. *Journal of the American Society for Information Science* 46: 400-25.

Holloway, Todd et al. 2005. Analyzing and visualizing the semantic coverage of Wikipedia and its authors: research articles. *Complexity* 12n3: 30-40. Preprint available online at URL: http://arxiv.org/ftp/cs/papers/0512/0512085.pdf.

How to classify works using ACM's Computing classification system. ACM website. Available online at URL: http://www.acm.org/about/class/how-to-use.

Kingston, John. 2002. Ontology, knowledge management, knowledge engineering and the ACM classification scheme. In *Proceedings of ES'02, the 22nd Annual International Conference of the British Computer Society's Specialist Group on Artificial In-*

telligence, Cambridge,(Great Britain), 10-12 December 2002. Available online at URL: http://www.inf.ed.ac.uk/publications/online/0169.pdf.

Leydesdorff, Loet. 1987. Various methods for the mapping of science. *Scientometrics* 11: 291-320.

Marshakova, Irina V. 1973. A system of document connection based on references. *Scientific and technical information serial of VINITI* 6n2: 3-8.

McCain, Katherine W. 1998. Neural networks research in context: a longitudinal journal cocitation analysis of an emerging interdisciplinary field. *Scientometrics* 41: 389-410.

Mirkin, Boris et al. 2008. Representing a computer science research organization on the ACM Computing classification system. In Eklund, Peter, Haemmerlé, Ollivier, *Supplementary proceedings of the 16th International Conference on Conceptual Structures (ICCS-2008), CEUR Workshop Proceedings, 354, RWTH Aachen University*. Available online at URL: http://ftp.informatik.rwth-aachen.de/ Publications/CEUR-WS/Vol-354/p19.pdf.

Moed, Henk F. and Visser, Martin S. 2007. *Developing bibliometric indicators of research performance in computer science*. Research report to the Council for Physical Sciences of the Netherlands Organisation for Scientific Research (NWO), Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands. Available online at URL: http://www.cwts.nl/pdf/ NWO_Inf_Final_Report_V_210207.pdf

Moya-Anegón, Félix et al. 2004. A new technique for building maps of large scientific domains based on the cocitation of classes and categories, *Scientometrics* 61: 129-45.

Osinska, Veslava. 2005. Szczegółowa klasyfikacja przedmiotowa Nauk Komputerowych w kontekście zawartości zasobów sieciowych (Computer Science Classification in the context of WEB resources). In *VIII Forum Polskie Towarzystwo Informacji Naukowej (Polish Society of Information Science), Zakopane (Poland) 12-14 October 2005*. PTIN 6.

Osinska, 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.

Osinska, 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.

Samoylenko, I., et al. 2006. Visualizing the scientific world and its evolution. In: *Journal of the American Society for Information Science and Technology* 57: 1461-69.

Scharnhorst, Andrea. 2003. Complex networks and the Web: insights nonlinear physics. *Journal of Computer-Mediated Communication* 8n4. Available online at URL: http://jcmc.indiana.edu/vol8/issue4/scharnhorst.html.

Small, Henry. 1973. Co-citation in the scientific literature: a new measurement of the relationship between two documents. *Journal of the American Society for Information Science* 24: 265-69.

Smiraglia, Richard P. 2008. ISKO 10's bookshelf. *Knowledge organization* 35:187-91.

Sosinska-Kalata, Barbara. 2002. Classification structures as web resources organization. In *MISSI 2002: III Polish Conference multimedia and Network Information Systems, Wroclaw (Poland), 19-20 September*. Available online at URL: http://www.zsi.pwr.wroc.pl/zsi/missi2002/pdf/s403.pdf (in polish).

Stasko, John T. et al. 2002. An evaluation of space-filling information visualizations for depicting hierarchical structures. *International journal of human-computer studies* 53: 663-94.

Syslo, Maciej M. and Kwiatkowska, Anna B. 2005. Informatics versus information technology—how much informatics is needed to use information technology—a school perspective. In Mittermeir, R.T., ed., *From computer literacy to informatics fundamentals*, *Proceedings of the International Conference on Informatics in Secondary Schools, Evolution and Perspectives*, ISSEP, LNCS 3422, pp. 178-88.

Umpleby, Stuart. 2000. *Defining 'cybernetics'*. Available online at URL: http://www.asc-cybernetics.org/index.htm.

Ware, Colin. 2004. *Information visualization: perception for design*, 2nd ed. San Francisco: Morgan Kaufmann.

White, Howard D. and McCain, Katherine W. 1997. Visualization of literatures. In Williams, M.E. ed., *Annual review of information science and technology*, v. 32. Medford, NJ: Information Today, pp. 99-168.

Yang, Christopher C. et al. 2003. Visualization of large category map for internet browsing. *Decision support systems* 35: 89-102.

Zeller, Daniel. 2007. Hypothetical model of the evolution and structure of science. New York, NY. Courtesy of Daniel Zeller. In Katy Börner and Julie M. Davis, Julie M. eds., 3rd Iteration (2007): The Power of Forecasts, *Places and Spaces: Mapping Science*. Available online at URL: http://www.scimaps.org/maps/map/hypothetical_model_o_51/.