This is the table of contents of Katy Börner, Chaomei Chen, & Kevin Boyack: Visualizing Knowledge Domains. Annual Review of Information Science & Technology, Volume 37, 2003. (in press)

Visualizing Knowledge Domains[1]

Katy Börner
School of Library and Information Science, Indiana University, Bloomington, IN 47405, USA
katy@indiana.edu

Chaomei Chen
College of Information Science and Technology, Drexel University, Philadelphia, PA 19104, USA
chaomei.chen@cis.drexel.edu

Kevin W. Boyack
Sandia National Laboratories, Albuquerque, NM 87185, USA
kboyack@sandia.gov

"The purpose of computing is insight – not numbers."

R. W. Hamming (1962)

[1] All figures in this chapter will be available in color at http://ella.slis.indiana.edu/~katy/arist02.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

ABSTRACT

This chapter reviews visualization techniques that can not only be utilized to map the ever-growing domain structure of scientific disciplines but that also support information retrieval and classification. In contrast to the comprehensive surveys done in a traditional way by Howard White and Katherine McCain (1997; 1998), the current survey not only reviews emerging techniques in interactive data analysis and information visualization, but also visualizes bibliographical structures of the field as an integral part of our methodology. The chapter starts with a review of the history of knowledge domain visualizations. We then introduce a general process flow for the visualization of knowledge domains and explain commonly used techniques. In the interest of visualizing the domain this article reviews, we introduce a bibliographic data set of considerable size, which includes articles from the citation analysis, bibliometrics, semantics, and visualization literatures. Using a tutorial style, we then apply various algorithms to demonstrate the visualization effects produced by different approaches and compare the different visualization results. At the same time, the domain visualizations reveal the relationships within and between the four fields that together form the topic of this chapter, domain visualization. We conclude with a discussion of promising new avenues of research and a general discussion.

1 INTRODUCTION

Painting a big picture of scientific knowledge has always been desirable for various reasons. Traditional approaches are brute-force in nature – scholars have to sort through through the mountains of literature to conduct their surveys. Obviously, this is time-consuming, difficult to repeat, and subjective. The task is enormous in its complexity. Sifting through recently published documents to find ones that will later be recognized as important is labor-intensive. Traditional approaches are increasingly hard to keep up with the pace of information growth. When it comes to a multidisciplinary field of study, it is rather difficult to maintain an overview of what is going on. Painting the "big picture" of an ever-evolving scientific discipline has been akin to the situation described in a widely known Indian legend about blind men and an elephant. As the

legend goes, six blind men were trying to find out what an elephant looks like. They touched different parts of the elephant and quickly jumped to their conclusions. The one touching the body said it must be like a wall; the one touching the tail said it is like a snake; the one touching the legs said it is like a tree trunk; etc. But science does not stand still; the steady stream of new scientific literature creates a continuously changing structure. The resulting disappearance, fusion, and emergence of research areas ads another twist to the legend – it is as if the elephant is running and dynamically changing in shape.

Domain visualization is an emerging field of study that is in a similar situation. Relevant literature is spread across disciplines that traditionally have few connections. Researchers looking at the domain from a particular discipline cannot possibly have an adequate understanding of the whole. As noted by White and McCain (1997), the new generation of information scientists are on the one hand technically driven in the new rush of visualizing scientific disciplines. On the other hand, they are rather limited with regard to what has been done in terms of bridging between pioneers' theories and practices and today's more enabling technologies. If the difference between the past and present generations is in the power of available technologies, what they have in common is the ultimate goal – to reveal the development of scientific knowledge. Today's use of this knowledge has expanded to include studies of scholarly communities and networks, the growth and evolution of fields, the diffusion of research topics, individual authors, or institutions, etc.

The survey of White and McCain (1997) was done in a traditional way, i.e., using manual and intellectual analysis. Since then the size and the scope of the field has exploded and it is now well beyond the reach of traditional survey methods. The types of enabling techniques needed to do current analyses quickly and effectively are precisely the ones that belong to the domain visualization toolkit. These new techniques allow us to streamline the practice with an unprecedented scalability and repeatability. To form the big picture itself is also a typical problem in domain visualization. How to choose the source of data, how to analyze and visualize the data, and how to make sense of what is in the picture, are decisions to be made by the new generation of information cartographers.

This work does not attempt to update the work of the former survey by providing an extensive bibliography with commentary on the field of literature mapping and visualization, but rather provides an overview of the many techniques and variations used in the process of mapping and visualizing knowledge domains. It also offers an opportunity to compare and contrast several different visualizations of the same data so as to illustrate characteristics of particular mapping techniques. This chapter does not cover some potentially relevant and important issues including user and task analysis, alternative input/output devices, visual perception principles, or evaluation of map relevance.

The balance of this covers the following information. Section 2 sketches the history of research on visualizing knowledge domains that is rooted in fields such as scientometrics, bibliometrics, citation analysis, and information visualization. Section 3 explains the general process flow of visualizing knowledge domains and the general structure of this chapter. We then review measures and approaches to determine bibliographic, linguistic, co-word, co-term, co-classification, content, or semantic similarities. Section 4 provides an overview of different

mathematical techniques that are commonly used to analyze and visualize bibliographic data. Appropriate visualization and interaction metaphors ensuring that resulting maps can be intuitively and effectively used are introduced as well. Section 5 introduces a bibliographic data set of considerable size that will be utilized in this paper to map research on the visualization of knowledge domains in section 6. We conclude with a discussion of promising new avenues of research and a general discussion.

2 HISTORY

Narin and Molls (1977) and White and McCain (1989) compiled the very first ARIST reviews of bibliometrics research. In (1997), White and McCain gave a vivid account of the history of citation analysis and its application to the visualization of literatures. Borgman in (1990; 2000) and her recent ARIST chapter with Furner (2002) gave a comprehensive overview of bibliometric methods that can be used to describe, explain, predict, and evaluate scholarly communication. Wilson's recent, very comprehensive review on informetrics covers bibliometrics research as well as other metric studies (2001). An in-depth account of theories and practices in the endeavor of mapping scientific frontiers is also the central topic of a forthcoming book (Chen, 2002).

2.1 Scientometrics, Bibliometrics, and Citation Analysis

Today's wide availability of citation index databases originated in the 1950s. Indexing in the 1950s was inconsistent and uncoordinated. There was widespread dissatisfaction with the array of traditional discipline-oriented indexing and abstracting services (Garfield, 1955). Eugene Garfield's pioneering paper in *Science* (Garfield, 1955) laid down the foundation of citation analysis today. In the words of White and McCain (1998):

> *"Eugene Garfield, the founder of ISI, devoted years to fulfilling his dream of creating a multidisciplinary citation index. The development of the Science Citation Index represented a fundamental breakthrough in scientific information retrieval. What began as a commercial product—a unique resource for scientists, scholars, and researchers in virtually every field of intellectual endeavor—has evolved into a sophisticated set of conceptual tools for understanding the dynamics of science. The concept of citation analysis today forms the basis of much of what is known variously as scientometrics, bibliometrics, infometrics, cybermetrics, and webometrics. Garfield's invention continues to have a profound impact on the way we think about and study scholarly communication."*

One of the pioneering domain visualization studies based on citation data is the creation of the historical map of research in DNA, which was done manually almost 40 years ago in early 1960s (Garfield, Sher, & Torpie, 1964). Soon thereafter Derek Price studied the same data in his classic work of mapping scientific networks (Price, 1961, 1963; Price, 1965). In domain visualization, interrelationships between research fronts are represented through spatial representations. Such spatial representations allow users to navigate the scientific literature based on the spatial patterns depicted.

Domain visualization aims to reveal realms of scientific communication as reflected through scientific literature and citation paths interwoven by individual scientists in their publications. There is indeed a profound connection between domain visualization and what Hjørland (1997) called domain analysis. Domain visualization can provide enabling techniques needed for domain analysis, especially in multidisciplinary and fast-moving knowledge domains. The field of domain visualization is also called scientography (Garfield, 1994), although the term scientography does not seem to be widely used.

Garfield (1994) also introduced the concept of longitudinal mapping. In longitudinal mapping, a series of chronologically sequential maps can be used to detect the advances of scientific knowledge. Analysts and domain experts can use longitudinal maps to forecast emerging trends for a subject domain. Since domain visualizations typically reference key works in a field, they are a good tool to enable the novice to become familiar with a field through easy identification of landmark articles and books, as well as members of the invisible college or specialties. The Web of Knowledge, released in 2000 to commemorate Dr. Eugene Garfield's 75th birthday, comprehensively addresses the history, theory, and practical applications of citation indexing and analysis (Cronin & Atkins, 2000).

Scientometrics is a distinct discipline that has emerged from citation-based domain visualization. Scientometrics is the quantitative study of scientific communications, which applies bibliometrics to scientific literature. Robert Merton and Eugene Garfield regard the late Derek De Solla Price (1922-1983) as the "father of scientometrics." Price made profound contributions to information science through his seminal work on networks of scientific papers (Price, 1965) as well as his landmark work *Little Science, Big Science, and Beyond* (Price, 1986).

In 1981, the Institute for Scientific Information (ISI) published the pioneering *Atlas of Science in Biochemistry and Molecular Biology* (1981). The Atlas was constructed based on a co-citation index associated with publications in the field over a one-year period. It featured 102 distinct clusters of articles. These clusters, representing research front specialties, form a snapshot of significant research activities in biochemistry and molecular biology. The construction of this pioneering Atlas took several months. Garfield and Small (1989) explained the role of citation structures in identifying the changing frontiers of science.

More recently, ISI has developed the SCI-Map software, which enables users to navigate a citation network. It has been used in numerous subject domains, including physics, chemistry, quantum systems, and other fields. For example, in 1994, Henry Small used SCI-Map to map AIDS research (Small, 1994). SCI-Map creates maps of individual research areas specified by the user. Given an author, paper, or keyword as a starting point, one can seed a map and then grow the map by specifying various desired connections at different thresholds of co-citation strength or distance. The network of connected nodes is formed by a series of iterations of clustering, including additional core papers with each successive node. The nodes are selected according to the strength of their links, and positioning is determined by the *geometric triangulation method* (see section 4.3.1).

In his most recent work, Small explored the notion of a passage through science (Small, 1999a, 2000). Passages linking the literature of different disciplines are likely to import or export a

method established in one discipline into another. This has been known as cross-disciplinary fertilization. As Small has noted, this reaching out or stretching can import or export methods, ideas, models, or empirical results from the author's field to the other field. This requires scientists to have not only a broad awareness of literature, but also the creative imagination to foresee how the outside information fits with the problem at hand. He developed algorithms to blaze a magnificent trail of more than 300 articles across the literatures of different scientific disciplines.

When speaking of citation indexing, one must also consider the Web-based citation database system ResearchIndex (formerly CiteSeer) developed by researchers at NEC Research Institute (Lawrence, Giles, & Bollacker, 1999). ResearchIndex allows users to search for various citation details of scientific documents available on the Web. This service provides a valuable complementary resource to ISI's citation databases. ResearchIndex takes advantage of being able to access full text versions of scientific documents on the Web by introducing a functionality called *citation context*. Not only can users search the database on various bibliographic attributes of a document such as the author, article title, and journal title, but they can also use the *citation context* function to retrieve a list of highlights as excerpts from citing documents and access detailed statements of its perceived value. This function provides an invaluable tool for researchers to judge the nature of an influential article.

Typically, the act of referencing another author's work in a scholarly or research paper is assumed to reflect a direct semantic relationship between the citing and cited works. However, a macroanalysis of cited and citing documents in terms of broad subject dispersion and a microanalysis that examined the subject relationship between citing and cited documents presented by (Harter, Nisonger, & Weng, 1993; Nisonger, Harter, & Weng, 1992) suggest that the subject similarity among pairs of cited and citing documents is typically very small, supporting a subjective, psychological view of relevance and a trial-and-error, heuristic understanding of the information search and research processes.

The notion of bibliometric mapping has been further developed by researchers in the Netherlands, in particular Noyons and van Raan (Noyons, Moed, & Luwel, 1999; Noyons & Van Raan, 1998; van Raan, 2000). Noyons and van Raan have developed special mathematical techniques for bibliometric mapping. The basic assumption is that each research field can be characterized by a list of the most important keywords. Each publication in the field can in turn be characterized by a sub-list of these global keywords. Such sub-lists are like DNA fingerprints of these published articles. By matching keyword-based fingerprints, one can measure the similarity between a pair of publications. The more keywords two documents have in common, the more similar the two publications are, and the more likely they come from the same research area or research specialty at a higher level. Following the DNA metaphor, if two publications' fingerprints are similar enough, they are bound to come from the same species. In (Noyons, Moed, & Luwel, 1999), they incorporate performance assessment into the creation of bibliometric maps in order to measure the impact level of different sub-fields and themes and to address strategic questions such as: who is where in the subject domain, and how strong is their research?

2.2 Map Generation and Visualization

In 1987, the National Science Foundation (NSF) panel report (McCormick, DeFanti, & Brown) recommended that NSF fund immediate and longer-term research in what is now known as the field of scientific visualization. At that time, there were about 200 supercomputers in the United States. These supercomputers generated a vast amount of numerical data, mainly through computationally intensive simulation of physical processes. Virtual wind tunnels and high-resolution predictive weather models are typical examples of scientific calculations that require visualization to present their output in an understandable form.

Scientific visualizations map physical phenomena onto 2D or 3D representations that are typically not very interactive. In contrast, Information Visualization (IV) aims at an interactive visualization of abstract non-spatial phenomena such as bibliographic data sets, web access patterns, etc.

Advances of information visualization were significantly driven by information retrieval research. A central problem for information retrieval researchers and practitioners is how to improve the efficiency and effectiveness of information retrieval. Generally speaking, the more a user knows about her search space, the more likely that her search will become more effective. Many information visualization systems depict the overall semantic structure of a collection of documents. Users can use this structural visualization as the basis for their subsequent browsing and search. Card (1996) and Hearst (1999) gave surveys of visualizing retrieval results.

Edward Tufte has published three seminal books (Tufte, 1983; 1990, 1997) on display and visualization. Although his 1983 and 1990 books were published prior to the emergence of information visualization as a distinct field, they are highly regarded in the information visualization community. In particular, Tufte's in-depth case study on the disaster of the launch of the space shuttle Challenger is a though-provoking example.

Research in hypertext started to emerge as a distinct field of study in late 1980s following a number of pioneering hypertext systems, notably HyperCard from Apple and NoteCards from Xerox PARC (Halasz, 1988; Halasz, Moran, & Trigg, 1986). A core issue of hypertext research is to enable users to easily navigate in hypertext spaces (Conklin, 1987). The thinking-by-association tradition of hypertext and the World-Wide Web later on has been widely attributed to the visionary Memex envisaged by Vannevar Bush (Bush, 1945). Researchers have studied a variety of navigation cues to help users to move around. One of the most popular recommendations for designers of hypertext systems is to have an overview map of the entire hypertext structure (Halasz, 1988). Advances have been made in automatically generating overview maps that can help users navigate. The mid-1990s saw a wide spread use of the World-Wide Web. The sheer size of the Web has posed an unprecedented challenge for mapping.

Geographic Information Systems (GIS) represent a gray area between information visualization and traditional cartography. Geographic coordinates provide a most convenient and natural organizing framework. A geographic framework can accommodate a wide variety of information. Thematic maps provide a rich metaphor for a class of information visualization known as information landscape. Notable examples include SPIRE/Themescape (Wise et al.,

1995) and BEAD (Chalmers, 1992). A recent book by Martin Dodge and Rob Kitchin (2000) is a good source of examples of how geography influences mapping cyberspace.

The number of review and survey articles on information visualization is steadily increasing (Card, 1996; Hearst, 1999; Herman, Melançon, & Marshall, 2000; Hollan, Bederson, & Helfman, 1997; Mukherjea, 1999). The first edited volume (Card, Mackinlay, & Shneiderman, 1999) and the first authored monograph (Chen, 1999a) both appeared in 1999. There are currently about a half dozen books on the market on information visualization (Card et al., 1999; Chen, 1999a; Spence, 2000, 2001; Ware, 2000), as well as a related book on algorithms for graph visualization (Battista, Eades, Tamassia, & Tollis, 1999). A new, peer-reviewed international journal *Information Visualization* is to be launched in March 2002 by Palgrave, Macmillan's global academic publishing. Today, there are numerous workshops and special issues held all over the world relating to information visualization.

Journals such as the *Journal of the American Society for Information Science and Technology* (JASIST) and *Scientometrics* have provided the focal forum for domain visualization. These journals traditionally have their main readership in Library and Information Science (LIS), rather than from other potentially relevant disciplines such as computer science, information visualization, and geographic information systems. In the past 15 years, information retrieval has enjoyed its prominent position in the mainstream information visualization research, but other research areas such as citation analysis and domain analysis remain tied to a relatively focused scientific community.

Major information visualization and interaction design techniques as they pertain to the visualization of knowledge domains are discussed in section 4.4.

3 PROCESS FLOW OF VISUALIZING KNOWLEDGE DOMAINS

White and McCain (1997) defined five models of literatures: (1) bibliographic, (2) editorial, (3) bibliometric, (4) user, and (5) synthetic. With the computerized tools and techniques available today, the lines between these traditional models can become blurred. The model used by many researchers today might be described as a *USER META MODEL*. It is first a *user model* in that it is a reduction of the literature based on a user's searches, queries, profiles, or filters, often generated quickly from computerized access to literature data sources, and formulated to provide answers to specific questions. It fulfills the role of the *bibliographic* or *meta model* in that it contains metadata – authors, titles, descriptive terms, dates, etc. – that can be used to define relationships pertinent to mapping, and also to display attributes of the data in modern visualizations. These data also often contain, or can easily be used to generate, *bibliometric* data – citation counts, term distributions, attributes by year, impact factors, etc. – that can be easily displayed by visualizations and that enhance map interpretation. Bibliometric attributes also allow for thresholds and rankings, which can be used to limit data to those deemed most pertinent or important by the user.

The *user meta model* is closely related to the process by which domain maps or visualizations are produced. An overview of this process, with many of its possible perturbations, is shown in Figure 1. The general steps in this sequence are (1) data extraction, (2) definition of unit of analysis, (3) selection of measures, (4) calculation of a similarity between units, (5) ordination,

or the assignment of coordinates to each unit, and (6) use of the resulting visualization for analysis and interpretation. Steps four and five of this process are often distilled into one operation, which can be described as data layout.

DATA EXTRACTION	UNIT OF ANALYSIS	MEASURES	LAYOUT (often one code does both similarity and ordination steps)		DISPLAY
			SIMILARITY	ORDINATION	
SEARCHES ISI INSPEC Eng Index Medline ResearchIndex Patents etc. BROADENING By citation By terms	COMMON CHOICES Journal Document Author Term	COUNTS/FREQUENCIES Attributes (e.g. terms) Author citations Co-citations By year THRESHOLDS By counts	SCALAR (unit by unit matrix) Direct citation Co-citation Combined linkage Co-word / co-term Co-classification VECTOR (unit by attribute matrix) Vector space model (words/terms) Latent Semantic Analysis (words/terms) incl. Singular Value Decomp (SVD) CORRELATION (if desired) Pearson's R on any of above	DIMENSIONALITY REDUCTION Eigenvector/ Eigenvalue solutions Factor Analysis (FA) and Principal Components Analysis (PCA) Multi-dimensional scaling (MDS) Pathfinder networks (PFNet) Self-organizing maps (SOM) includes SOM, ET-maps, etc. CLUSTER ANALYSIS SCALAR Triangulation Force-directed placement (FDP)	INTERACTION Browse Pan Zoom Filter Query Detail on demand ANALYSIS

Figure 1. Process flow for mapping knowledge domains.

The next few sections of this paper will address the process described in Figure 1. The balance of section 3 is used to review units of analysis, measures, and simple approaches to determine appropriate similarities between units.

Section 4 is designed to address both the Figure 1 process and two major problems in communicating information: (1) multivariate data need to be displayed on the two-dimensional surface of either paper or computer screen and (2) large amounts of data must be displayed in a limited space with limited resolution. The first problem is tackled by applying mathematical dimensionality reduction algorithms to map n-dimensional data into a 2-D or 3-D space. The purpose of these algorithms is to place objects that are similar to one another in n-dimensions close to each other and to place dissimilar objects far apart. This process is also called *ordination*. Cluster techniques can be used to further group similar objects together. Commonly used techniques are presented in section 4. The second problem is typically minimized by applying interaction (panning, filtering) and distortion techniques (fisheye) as discussed in section 4.4.

The general consensus in relevant fields such as information visualization and geographic cartography is that multiple maps are preferred to a single map whenever possible. This is because each map may show different insights from the same data set. Therefore, section 5 introduces a bibliographic data set of the subject domain that will be utilized to demonstrate the different similarity measures, data mining techniques, and visualization approaches. Section 6 shows multiple maps and comparisons of the example bibliographic data set, focusing on the key issues and key components uncovered through this multi-perspective approach.

3.1 Units of Analysis

The first step in any mapping process is the extraction of appropriate data, as will be exemplified in section 7. We will not deal further with the extraction issue, search strategies, or the like, but simply note that the quality of any mapping or visualization is necessarily constrained by the quality of underlying data. The number of documents retrieved to generate a domain map can range from several hundred to tens of thousands.

Selection of a unit of analysis, relevant to the questions one desires to answer, is the second step. The most common units in the mapping of literatures are journals, documents, authors, and descriptive terms or words. Each presents different facets of a domain and enables different types of analysis. For instance, a map of journals can be used to obtain a macro view of science (Bassecoulard & Zitt, 1999), showing the relative positions and relationships between major disciplines. Journal maps are also used on a much smaller scale (Ding, Chowdhury, & Foo, 2000; Leydesdorff, 1994; McCain, 1998) to show fine distinctions within a discipline.

Documents (articles, patents, etc.) are the most common unit used to map or visualize a knowledge domain. These maps are used for a variety of purposes, including document retrieval, domain analysis (Small, 1999a, 2000), informing policy decisions, or assessing research performance (Noyons, 2001; Noyons, Moed, & Luwel, 1999; Noyons, Moed, & Van Raan, 1999; Noyons & Van Raan, 1998; Noyons & van Raan, 1998), and science and technology management or competitive intelligence (Boyack, Wylie, & Davidson, 2002).

Author-based maps are also relatively common and occur in two main forms. Author co-citation maps (Chen, 1999b; Chen, Paul, & O'Keefe, 2001; Ding, Chowdhury, & Foo, 1999; Lin & Kaid, 2000; White & McCain, 1998) are typically used to infer the intellectual structure of a field. By contrast, co-authorship maps are used to show the social network of a discipline or department (Mahlck & Persson, 2000). Co-authorship maps have been used by Glanzel and co-workers at the Hungarian Academy of Sciences for a series of studies designed to reveal international collaborations (Glänzel, 2001; Glänzel & DeLange, 1997). Newman has studied the structure of scientific networks from a statistical point of view (Newman, 2001a, 2001b). His techniques, while not done from a mapping or visualization perspective, are relevant and scale to very large systems (e.g., 1.5 million authors from Medline).

Semantic maps, often known as co-word analyses, are used to understand the cognitive structure of a field (Bhattacharya & Basu, 1998; Cahlik, 2000; DeLooze & Lemarie, 1997; He, 1999; Salvador & Lopez-Martinez, 2000). These are generated from different textual sources including single words extracted from titles of articles, descriptive terms, or publisher-assigned descriptors supplied by a database vendor (e.g., ISI keywords). Earlier maps were enabled by the popularization and use of the Leximappe software (Callon, Courtial, Turner, & Bauin, 1983). However, Leximappe never really reached the US – researchers here tended to use standard bibliographic retrieval software (e.g., Dialog searching or Word Start) to collect co-descriptor or co-classification data. Ron Kostoff at the Office of Naval Research even wrote his own programs for co-word extraction and analysis. Many users today have computational tools[2] that allow them to do their own term extraction and mapping. Kostoff and coworkers have developed the

[2] For example, WordStat, available at http://www.simstat.com/wordstat.htm

Database Tomography technique, which they use to create science and technology roadmaps (Kostoff, Eberhart, & Toothman, 1998). Some confusion can be caused by reference to "semantic space" – which most often refers not to co-word maps, but rather to document maps using words or terms as labeling features.

There is no explicit reason why multiple units (e.g., journals and authors) cannot be used in the same map, but it is not commonly done. One example of multiple units is the work by White and Griffith (1982), an author co-citation study in which useful phrases were used to "self-label" factors.

3.2 Measures and Similarity Calculation

3.2.1 Measures

Measures have been defined very succinctly by White and McCain (1997), and rather than muddy the waters, we simply choose to quote their work here for completeness:

> *"We use certain technical terms such as intercitation, interdocument, co-assignment, co-classification, co-citation, and co-word. The prefix 'inter-' implies relationships between documents [or units]. The prefix 'co-' implies joint occurrences within a single document [or unit]. Thus, intercitation data for journals are counts of the times that any journal cites any other journal, as well as itself, in a matrix. (The citations appear in articles, of course.) The converse is the number of times any journal is cited by any other journal. The same sort of matrix can be formed with authors replacing journals. Interdocument similarity can be measured by counting indicators of content that two different documents have in common, such as descriptors or references to other writings (the latter is known as bibliographic coupling strength). Co-assignment means the assignment of two indexing terms to the same document by an indexer (the terms themselves might be called co-terms, co-descriptors, or co-classifications). Co-citation occurs when any two works appear in the references of a third work. The authors of the two co-cited works are co-cited authors. If the co-cited works appeared in two different journals, the latter are co-cited journals. Co-words are words that appear together in some piece of natural language, such as a title or abstract. Bother 'inter-' and 'co-' relationships are explicit and potentially countable by computer. Thus, both might yield raw data for visualization of literatures."*

To this we add a few definitions. A "citation" is the referencing of a document by a more recently published document. The document doing the citing is the "citing" document, and the one receiving the citation is the "cited" document. Citations may be counted and used as a threshold (e.g., only keep the documents that have been cited more than 5 times) in a mapping exercise. Other terms used to describe citing and cited numbers are "in-degree" or the number of times cited, and "out-degree" or the number of items in a document's reference list. Journal impact factors calculated from citation counts are published by ISI, and can be used to enhance

visualizations, as can the raw citation counts themselves.[3] An excellent review of bibliometric and other science indicators was also provided by King (1987).

3.2.2 Simple Similarities

Similarity between units is typically based on one of the following.

Citation linkages – These include direct citation linkage, co-citation linkage[4], bibliographic coupling (Kessler, 1963), longitudinal coupling (Small, 1995), and Small's combined linkage method (Small, 1997). Citation linkage similarities are naturally constrained to use with data derived from citation databases, such as the Science Citation Index or a patent database.

Co-occurrence similarities – The most common co-occurrence similarities include co-term, co-classification, author co-citation, and paper co-citation. Two of the more common similarity formulas used with co-occurrence are the simple cosine and Jaccard indices. Each counts the number of attributes common between two units (e.g., the number of terms in common between two articles), but differ in their normalization. Chen and Lynch (1992) developed an asymmetric cluster function and showed that it better represents term associations than the popular cosine function. Chung and Lee (2001) recently compared six different co-term association measures and discuss their relative behavior on a set of documents. Rorvig (1999) explored multiple similarity measurements on TREC document sets, finding that cosine and overlap measures best preserved relationships between documents. The *co-word similarity* uses the same types of associations as the co-term, but is commonly based on words extracted from the titles and/or abstracts of articles by counting the number of times any two words appear in the same text segment.

3.2.3 Vector Space Model

The Vector Space Model (VSM) was developed by Gerald Salton (Salton, Yang, & Wong, 1975). It is an influential and powerful framework for storing, analyzing, and structuring documents. Originally developed for information retrieval, the model is a widely used framework for indexing documents based on term frequencies. Its three stages are document indexing, term weighting, and computation of similarity coefficients:

- *Document indexing:* Each document (or query) is represented as a vector in a high dimensional space. Dimensionality is determined by the number of unique terms in a document corpus. Non-significant words are removed from the document vector. A stop list, which holds common words, is used to remove high frequency words.[5]
- *Term weighting:* Terms are weighted to indicate their importance for document representation. Most of the weighting schemes such as the inverse document

[3] Nederhof and Zwaan (1991) collected peer judgments on the quality of journals by means of a world-wide mail survey among 385 scholars and probed the quality of the coverage by the SSCI and the AHCI of both core and noncore journals. Results showed convergence with those based on journal impact factors.

[4] Co-citation is known to have a low "recall" of clusters because only papers which have citation links within the data set and the defined time window can be classified. "Co-citation and bibliographical coupling offer cross sectional views given narrow, one year citing periods. Longitudinal coupling becomes effective only when wider periods are used. At the end of a time period documents will be linked through their references to earlier items, while at the beginning, linking will be through citations received." (Small, 1997, p.278-279).

[5] In general, 40-50% of the total number of words in a document is removed.

frequency (see below) assume that the importance of a term is proportional to the number of documents the term appears in. Long documents usually have a much larger term set than short documents, which makes long documents more likely to be retrieved than short documents. Therefore, document length normalization is employed.

- *Computation of similarity coefficients:* The similarity between any two documents (or between a query and a document) can subsequently be determined by the distance between vectors in a high-dimensional space. Word overlap indicates similarity. The most popular similarity measure is the cosine coefficient, which defines the similarity between two documents by the cosine of the angle between their two vectors. It resembles the inner product of the two vectors, normalized (divided) by the products of the vector lengths (square root of the sums of squares).

The discriminative power of a term is determined by the well-known $tf \times idf$ model, in which *tf* denotes the term frequency and *idf* represents the inverse document frequency. Each document can be represented by an array of terms T and each term is associated with a weight determined by the $tf \times idf$ model. In general, the weight of term T_k in document D_i, is estimated as follows:

$$w_{ik} = \frac{tf_{ik} \times log\left(\frac{N}{n_k}\right)}{\sqrt{\sum_{j=1}^{T} \left(tf_{ij}\right)^2 \times log\left(\frac{N}{n_j}\right)^2}},$$

where tf_{ik} is the number of occurrences of term T_k in D_i, N is the number of documents in a given collection, and n_k represents the number of documents containing term T_k. The document similarity is computed as follows based on corresponding vectors $D_i = (w_{i1}, w_{i2}, ..., w_{iT})$ and $D_j = (w_{j1}, w_{j2}, ..., w_{jT})$:

$$sim_{ij}^{content} = \sum_{k=1}^{T} w_{ik} \times w_{jk} .$$

Document similarity can be used to group a large collection of documents into a number of smaller clusters such that documents within a cluster are more similar than documents in different clusters.

The vector space model provides an easy way to assess document similarities based on word matches. Note that different meaning of words – e.g., the bird “crane” and the “crane” on a construction site – cannot be detected. This is known as the "vocabulary mismatch problem" the solution of which requires methods that examine the context of words such as Latent Semantic Analysis (see section 4.1.2); Lexical Chaining, a notion derived from work in the area of textual cohesion in linguistics (Halliday & Hasan, 1976); or the automatic discovery of vocabulary and thesauri (Mostafa, Quiroga, & Palakal, 1998).

Different applications of the vector space model are presented in (Salton, Allan, & Buckley, 1994; Salton, Allan, Buckley, & Singhal, 1994; Salton & Buckley, 1988, 1991; Salton et al., 1975). For a critical analysis of the vector space model for information retrieval consult (Raghavan & Wong, 1986).

4 ENABLING TECHNOLOGIES

This section describes enabling techniques with regard to the analysis and visualization of knowledge. In particular, we describe methods that are generally used to create (interactive) visualizations of knowledge domains.

Subsection 4.1 will introduce dimensionality reduction techniques that can be applied to represent n-dimensional data by a small number of salient dimensions and thus to display multivariate data on the two-dimensional surface of either paper or computer screen. Several of these algorithms produce a 2-D or 3-D spatial layout in which similar objects are close to one another. This process is also called *ordination*. Cluster analysis, presented in subsection 4.2, can be used to further group similar objects together, and to determine category boundaries and labels. Some of the algorithms presented in subsection 4.1 generate a document-by-document similarity matrix that can be visualized by spatial configuration algorithms, see subsection 4.3. Last but not least, subsection 4.4 presents the application of interaction and distortion techniques that aim to solve the second information communication problem – to display large amounts of data must be displayed in a limited space with limited resolution. For each technique we will give a general description, discuss its value for visualizing knowledge domains, and give references to further reading and code if available. We conclude with a general comparison of different techniques.

4.1 Dimensionality Reduction Techniques

Dimensionality reduction is an effective way to derive useful representations of high-dimensional data. This section reviews a range of techniques that have been used for dimensionality reduction, including Eigenvalue/Eigenvector decomposition, Factor Analysis (FA), Multidimensional Scaling (MDS), Pathfinder Network Scaling (PF), and Self-Organizing Maps (SOMs).[6]

4.1.1 Eigenvalue/Eigenvector Decomposition

Eigenvalue/Eigenvector decomposition is a technique that has been widely used in scientific computation. Given an $N \times N$ matrix $\boldsymbol{A}$, if there exist a vector $\boldsymbol{v}$ and a scalar value $\boldsymbol{\lambda}$ such that $\boldsymbol{Av} = \boldsymbol{\lambda v}$. The vector $\boldsymbol{v}$ is an eigenvector, and the scalar value $\boldsymbol{\lambda}$ is a corresponding eigenvalue. Eigenvalue/Eigenvector decomposition is commonly used to reduce the dimensionality of a high-dimensional space while its internal structure is preserved. A related technique is called *singular value decomposition* (SVD), which is used in Latent Semantic Analysis (see sction 4.1.4).

Given a collection of points in a high-dimensional space, the eigenvalues of the covariance matrix reveal the underlying dimensionality of the space. Eigenvector analysis techniques encompass Principal Components Analysis (see section 4.1.2) and Empirical Orthogonal Functional Analysis. Common features of these eigenvalue problems are (1) the number of

[6] Principal component analysis is used in Eigenvalue and factor analysis. However, Eigen solutions can give coordinates for each document, while a factor analysis doesn't.

eigenvalues required is small relative to the size of the matrices and (2) the matrix systems are often very sparse or structured.

Sandia's VxInsight has the option of using an eigenvalue solver (Davidson, Hendrickson, Johnson, Meyers, & Wylie, 1998). However, in practice this solution, while mathematically robust, does not necessarily place dissimilar objects far apart and does not tend to produce discrete clusters.

4.1.2 Factor Analysis and Principal Components Analysis

The term Factor Analysis (FA) was first introduced by Thurstone (1931). Factor analysis is a multivariate exploratory technique that can be used to examine a wide range of data sets. Primary applications of factor analytic techniques are: (1) to reduce the number of variables and (2) to detect structure in the relationships between variables, or to classify variables. Therefore, factor analysis is applied as a data reduction or structure detection method. Contrary to other methods such as LSA, the factors can often be interpreted.

A key method in factor analysis is Principal Component Analysis (PCA), which can transform a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. An advantage of using factor analysis over traditional clustering techniques is that it does not force each object into a cluster. Objects can be classified in multiple factors, thus preserving an important type of phenomenon: truly important work is often universal.

There are many excellent books on factor analysis such as (Basilevsky, 1994; Gorsuch, 1983; Harman, 1976). PCA has been routinely used by information scientists, especially in author co-citation analysis (Chen & Carr, 1999; McCain, 1990, 1995; Raghupathi & Nerur, 1999; White & McCain, 1998). PCA was also employed in SPIRE (Hetzler, Whitney, Martucci, & Thomas, 1998; Wise, 1999; Wise et al., 1995), using a context vector similar to those constructed in LSA.

4.1.3 Multidimensional Scaling

Multidimensional Scaling (MDS) attempts to find the structure in a set of proximity measures between objects (Kruskal, 1977). This is accomplished by solving a minimization problem such that the distances between points in the conceptual low-dimensional space match the given (dis)similarities as closely as possible.

The result is a least-squares representation of the objects in a lower (often 2-dimensional) space. The MDS procedure is as follows:

- All objects and their distances are determined.
- A goodness-of-fit measure called stress is maximized to produce a scatterplot of the objects in a low-dimensional space.
- The dimensions are interpreted, keeping in mind that the actual orientations of the axes from the MDS analysis are arbitrary, and can be rotated in any direction. In

addition, one can look for clusters of points or particular patterns and configurations (such as circles, manifolds, etc.).

The real value of MDS is that it can be used to analyze any kind of distance or similarity matrix. These similarities can represent people's ratings of similarities between documents, similarity between objects based on co-citations, etc. Due to computational requirements, only small data sets can be processed with MDS. Additionally, no relationship data (links) can be displayed. There are numerous texts available for further reading (Borg & Groenen, 1996; Joseph Kruskal, B., 1964; Joseph B. Kruskal, 1964; Kruskal & Wish, 1984). KYST is a flexible Fortran program developed by Kruskal, Young, and Seery for MDS which is available on the Internet.[7,8]

A long acknowledged major weakness of MDS is that there are no quick and fast rules to interpret the nature of the resulting dimensions. In addition, analysts often need more local details and more explicit representations of structures. An MDS configuration is limited in meeting these needs. The use of Pathfinder Network Scaling technique and Pathfinder networks provide users with additional local details and explicit representations of structures than MDS configurations (see section 4.1.5).

MDS has been one of the most widely used mapping techniques in information science, especially for document visualization (Chalmers, 1992), author co-citation analysis (White & McCain, 1998), document analysis (Hetzler et al., 1998), science mapping (Small, 1999b), and visualizing group memories[9] (McQuaid, Ong, Chen, & Nunamaker, 1999), and performance assessment (Noyons, Moed, & Van Raan, 1999) to name just a few.

Recently, nonlinear MDS approaches have been proposed that promise to handle larger data sets. Examples are the global geometric framework for nonlinear dimensionality reduction named Isomap[10] proposed in (Tenenbaum, de Silva, & Langford, 2000) and nonlinear dimensionality reduction by locally linear embedding proposed by (Roweis & Saul, 2000). Both techniques have not been applied to the visualization of knowledge domains yet.

4.1.4 Latent Semantic Analysis

Latent Semantic Analysis (LSA), also called Latent Semantic Indexing (LSI), was developed to resolve the so-called vocabulary mismatch problem (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Landauer, Foltz, & Laham, 1998). LSA handles *synonymy* (variability in human word choice) and *polysemy* (same word has often different meanings) by considering the context of words. It uses an advanced statistical technique, singular value decomposition (SVD), to extract latent terms. A latent term may correspond to a salient concept that may be described by several keywords, for example, the concept of human-computer interaction. The procedure is as follows:

- Representative samples of documents are converted to a matrix of title/authors/abstract words by articles. Cell entries are word frequencies in the title/authors/abstract of a given document.

[7] http://elib.zib.de/netlib/mds/kyst.f.
[8] http://elib.zib.de/netlib/mds/kyst2a_manual.txt
[9] See http://ai.bpa.arizona.edu/go/viz/mds.html for an online demo.
[10] http://isomap.stanford.edu/

- After an information theoretic weighting of cell entries, the matrix is submitted to singular value decomposition (SVD).
- SVD constructs an n-dimensional abstract semantic space in which each original word is presented as a vector.
- LSA's representation of a document is the average of the vectors of the words it contains independent of their order.

Construction of the SVD matrix is computationally expensive. There are also cases in which the matrix size cannot be reduced effectively. Yet, an effective dimensionality reduction helps to reduce noise and automatically organizes documents into a semantic structure more appropriate for information retrieval. This is a prime strength of LSA – once the matrix has been calculated, retrieval based on a user's query is very efficient. Relevant documents are retrieved, even if they did not literally contain the query words. The LSA matrix can also be used to calculate term-by-term or document-by-document similarities for use in other layout routines.

There are numerous LSA web resources including the Telcordia (formerly BellCore) LSI page[11], a web site at the University of Colorado[12], or the University of Tennessee[13]. SVDPACKC[14] (Version 1.0) developed by Michael Berry comprises four numerical (iterative) methods for computing the singular value decomposition of large sparse matrices using double precision ANSI Fortran-77. The General Text Parser[15] (GTP), developed by Howard, Tang, Berry, and Martin at the University of Tennessee, is an object-oriented (C++) integrated software package for creating data structures and encoding needed by information retrieval models.

LSA has been used in Generalized Similarity Analysis (Chen, 1997b, 1999b), StarWalker (Chen & Paul, 2001), and the *LVis - Digital Library Visualizer* (Katy Börner, 2000; Börner, Dillon, & Dolinsky, 2000) visualizations, among others. LSA has also been used by Porter and colleagues for technology forecasting (Zhu & Porter, 2002).

4.1.5 Pathfinder Network Scaling

Pathfinder Network Scaling is a structural and procedural modeling technique which extracts underlying patterns in proximity data and represents them spatially in a class of networks called Pathfinder Networks (PFnets) (Schvaneveldt, 1990). Pathfinder algorithms take estimates of the proximities between pairs of items as input and define a network representation of the items that preserves only the most important links. The resulting Pathfinder network consists of the items as nodes and a set of links (which may be either directed or undirected for symmetrical or non symmetrical proximity estimates) connecting pairs of the nodes. Software for Pathfinder Network Scaling is available for purchase.[16]

[11] http://lsi.research.telcordia.com/
[12] http://lsa.colorado.edu/
[13] http://www.cs.utk.edu/~lsi/
[14] http://www.netlib.org/svdpack/
[15] http://www.cs.utk.edu/~lsi/soft.html
[16] http://www.geocities.com/interlinkinc/home.html

The essential concept underlying Pathfinder networks is pairwise similarity. Similarities can be obtained based on a subjective estimation or a numerical computation. Pathfinder provides a more accurate representation of local relationships than techniques such as MDS.

The topology of a PFNET is determined by two parameters q and r and the corresponding network is denoted as PFNET(r,q). The q-parameter constrains the scope of minimum-cost paths to be considered. The r-parameter defines the Minkowski metric used for computing the distance of a path. The weight of a path with k links is determined by weights w_1, w_2, ..., w_k of each individual link as follows:

$$W(P) = \left[\sum_{i=1}^{k} w_i^r \right]^{\frac{1}{r}}$$

The q-parameter specifies that triangle inequalities must be satisfied for paths with $k \leq q$ links:

$$w_{n_i n_k} = \left[\sum_{i=1}^{k-1} w_{n_i n_k}^r \right]^{\frac{1}{r}} \quad \forall\ k \leq q$$

When a PFnet satisfies the following three conditions, the distance of a path is the same as the weight of the path:

1. The distance from a document to itself is zero.
2. The proximity matrix for the documents is symmetric; thus the distance is independent of direction.
3. The triangle inequality is satisfied for all paths with up to q links. If q is set to the total number of nodes less one, then the triangle inequality is universally satisfied over the entire network.

The number of links in a network can be reduced by increasing the value of the r or q parameter. The geodesic distance between two nodes in a network is the length of the minimum-cost path connecting the nodes. A minimum-cost network (MCN), PFnet(r=∞, q=n-1), has the least number of links.

AuthorLink and ConceptLink[17] developed by Xia Lin and colleagues enable to create interactive author co-citation analysis maps based on PFNet or Self Organizing Maps (White, Buzydlowsky, & Xia, 2000).

Pathfinder Network Scaling is used in Generalized Similarity Analysis (GSA) a generic framework for structuring and visualizing distributed information resources (Chen, 1997a, 1998a, 1998b, 1999a). The original version of the framework was designed to handle a number of intrinsic interrelationships in hypertext documents, namely hypertext linkage, content similarity, and browsing patterns. GSA is based on the notion of virtual link structures to organize its structural modeling and visualization functionality. Virtual link structures are in turn determined by similarity measurements defined between a variety of entity types, for example,

[17] http://cite.cis.drexel.edu/

document-to-document similarity, author-to-author similarity, and image-to-image similarity. Not only can one extend similarity measurements to new entity-entity relationships, but one can also integrate different similarity measurements to form a new network of entities. For example, interrelationships between hypertext documents can be defined based on a combination of hypertext connectivity, word-occurrence-based similarity, and traversal-based similarity. The generic framework of GSA led to several subsequent extensions to deal with a diverse range of data, including co-citation networks and image networks.

The use of Pathfinder networks in GSA reduces the excessive number of links in a typical proximity network and therefore improves the clarity of the graphical representations of such networks. The extensibility and flexibility of Pathfinder networks have been demonstrated in a series of studies along with a range of other techniques. Some recent examples include StarWalker for social navigation (Chen, Thomas, Cole, & Chennawasin, 1999), trailblazing the literature of hypertext (Chen & Carr, 1999), author co-citation analysis (Chen, 1999b), and visualizations of knowledge domains (Chen & Paul, 2001).

4.1.6 Self-Organizing Maps

One of the most profound contributions made by artificial neural networks to information visualization is the paradigm of self-organizing maps (SOMs) developed by Kohonen (Deboeck & Kohonen, 1998; Kaski, Honkela, Lagus, & Kohonen, 1998; Kohonen, 1985; Kohonen et al., 2000). During the learning phase, a self-organizing map algorithm iteratively modifies weight vectors to produce a typically 2-dimensional map in the output layer that will exhibit as best as possible the relationship of the input layer.

SOM maps appear to be one of the most promising algorithms for organizing large volumes of information. However, they have some significant deficiencies, many of which are discussed in (Kohonen, 1995). These deficiencies comprise the absence of a cost function, and the lack of a theoretical basis for choosing learning rate parameter schedules and neighborhood parameters to ensure topographic ordering. There are no general proofs of convergence, and the model does not define a probability density. A Self-Organizing Map Program is available from the Kohonen Neural Networks Research Centre, Helsinki University of Technology[18].

SOM maps have been used to map millions of documents from over 80 Usenet newsgroups[19] and to map the World-Wide Web. Xia Lin was the first to adopt the Kohonen SOM for information visualization (Lin, 1997; Lin, Soergel, & Marchionini, 1991) to document spaces. His Visual SiteMaps[20] visualized clusters of important concepts drawn from a database.

ET-Maps were developed in 1995 by Hsinchun Chen and his colleagues in the Artificial Intelligence (AI) Lab at the University of Arizona[21]. They constitute a scalable, multi-layered, graphical SOM approach to automatic categorization of large numbers of documents or web sites (Chen & Rada, 1996; Chen, Schuffels, & Orwig, 1996). The prototype was developed using the Yahoo! Entertainment sub-category (about 110,000 homepages); hence the name ET-Map.

[18] http://www.cis.hut.fi/research/som_pak/
[19] http://websom.hut.fi/websom
[20] http://faculty.cis.drexel.edu/sitemap/
[21] http://ai.bpa.arizona.edu/

ET-Maps are category maps that group documents that share many noun phrase terms together in a neighborhood on a 2-D map. Each colored region represents a unique topic that contains similar documents. The size of a subject region is related to the number of documents in that category such that more important topics (if importance can be correlated to counts) occupy larger regions. Neighborhood proximity is applied to plot "subject regions" that are closely related in terms of content, close to each other on the map. ET-Maps show an "up-button" view of an information space to provide the user with a sense of the organization of the information landscape, e.g., what is where, the location of clusters and hotspots, and what is related to what. ET-Maps are multi-layer maps, with sub-maps showing greater informational resolution through a finer degree of categorization.

Focus+context techniques[22] have been used to display a large SOM effectively within a limited screen area (Yang, Chen, & Hong, 1999). Usability studies indicate that users tend to get lost when browsing multi-level SOM maps and continued to prefer to use a conventional text-based alphabetic hierarchy (Chen, Houston, Sewell, & Schatz, 1998). Today, ET-Maps come with two panels. The left panel is a "Windows Explorer like" interface that presents an alphabetic display of the topic hierarchy generated, while the right panel is the graphical display of the SOM output. On the left panel, a user can click on any category of interest and the system displays its sub-categories beneath. At the same time, those sub-categories are also displayed on the right panel, where the spatial proximity equals the semantic proximity. In addition, colors are employed to indicate how many layers a user can go down within a certain category. A working demo of ET-Maps can be explored at the AI Lab's website.[23] ET-Maps and Cartographic SOM Maps are discussed further and exemplified in section 6.

Multi-SOMs are a multi-maps extension of SOMs. An automatic way of naming the clusters to divide the map into logical areas, and a map generalization mechanism are introduced by (Polanco, Francois, & Lamirel, 2001), who also discuss the application potential of Multi-SOMs for visualization, exploration or browsing, and scientific and technical information analysis.

4.2 Cluster Analysis

The term *cluster analysis* (CA) was first used by Tryon (1939). Cluster analysis encompasses a number of different classification algorithms that aim to organize a "mountain" of information into manageable, meaningful piles, called clusters.

A clustering problem can be defined by a set of objects (e.g., documents) and a vague description of a set A. The goal of clustering is to divide the object set into objects belonging to A and a second set not in A. In this clustering problem, one first needs to determine what features are relevant in describing objects in A (intra-cluster similarity) and second, what features distinguish objects in A from objects not belonging to A (inter-cluster similarity).

Alternatively, a cluster problem can be formulated by a set of objects and a similarity or distance function. Here, the goal is to divide the object set into number of sub-sets (clusters) that best

[22] See section 4.4.3.
[23] http://ai3.bpa.arizona.edu/ent/entertain1/

reveal the structure of the object set. These can take the form of partitions or a hierarchically organized taxonomy.

Clusters should be highly internally homogenous (members are similar to one another) and highly externally heterogeneous (members are not like members of other clusters). Thus, the aim is to maximize intra-cluster similarity and minimize inter-cluster similarity. This can be formulated in terms of a *utility measure* that contrasts the sum of within-cluster similarities wSim by the sum of between-cluster similarities bSim:

$$\text{utility} = \text{wSim} / (\text{wSim} + \text{bSim}).$$

Given alternative partitions the one that shows the highest utility is selected.

Clustering algorithms can be distinguished based on a number of features such as unsupervised or supervised, divisive or agglomerative, incremental or non-incremental, deterministic or non-deterministic, hierarchical or partitioning, iterative or non-iterative, single link, grouped average, or complete link clustering. Interestingly, no clustering algorithm has been shown to be particularly better than others when producing the same number of clusters (Hearst, 1999). However, experience demonstrates that some choices seem to fit some kinds of data better than others (e.g., correlation and complete linkage works very well for our ACA/JCA data) and there have been "bakeoffs" between clustering approaches (comparing single link, complete link, Ward's trace, centroid, etc.) that suggest that some approaches are more "reliable" than others for generic data sets. An excellent review of clustering algorithms can be found in (Han & Kamber, 2000).

In IV, clustering techniques are frequently applied to group semantically similar objects so that object set boundaries can be presented. The automatic assignment of cluster labels is yet another topic of high relevance for information visualization. For example, work by (Pirolli, Schank, Hearst, & Diehl, 1996) automatically computes summaries of the contents of clusters of similar documents providing a method for navigating through these summaries at different levels of granularity.

4.3 Spatial Configuration

Attributes of a data set can often be cast in the form of a similarity or distance matrix. Ordination techniques such as triangulation or force directed placement take a set of documents, their similarities/distances, and parameters and generate a typically 2-dimensional layout that places similar documents closer together and dissimilar ones further apart.

4.3.1 Triangulation

Triangulation is an ordination technique that maps points from an n-dimensional space into a typically two-dimensional one (Lee, Slagle, & Blum, 1977). It starts by placing a randomly selected point at the origin of the coordinate system. Next, the most similar object is determined and the second object is placed at a specified distance from the first object. The location of the third object is defined by the distance to the subsequent two objects (triangulation). Subsequently, the notion of repulsion from the origin is used to select the quadratic solution furthest from the origin – the spatial layout grows outwards.

Compared to classical ordination methods, triangulation is computationally inexpensive. The resulting layouts exactly represent the distances between single data points but lack global optimization.

Triangulation was used by Henry Small in the context of information visualization (Small, 1999b). His *Map of Science* is a series of nested maps showing the multi-dimensional landscape of science at five levels of aggregation.

4.3.2 Force Directed Placement

Force Directed Placement (FDP) can be used to sort randomly placed objects into a desirable layout that satisfies the given similarity relations among objects as well as aesthetics for visual presentation (symmetry, non-overlapping, minimized edge crossings, etc.) (Battista, Eades, Tamassia, & Tollis, 1994; Fruchterman & Reingold, 1991). FDP views nodes as physical bodies and edges as springs (or weighted arcs) connected to the nodes providing forces between them. Nodes move according to the forces on them until a local energy minimum is achieved. In addition to the imaginary springs, other forces can be added to the system in order to produce different effects. Many visual examples of these force models can be found in (Battista et al., 1994).

The FDP method is easy to understand and implement. However, it can be very slow for large graphs – in each iteration step the forces between all nodes have to be computed and considered to optimize the spatial layout. Modifications to a traditional force-directed approach have been made in the VxInsight ordination algorithm (Davidson, Wylie, & Boyack, 2001), VxOrd, and have resulted in a dramatic increase in computational speed. VxOrd accepts pairwise scalar similarity values as the arc weights, employs barrier jumping to avoid trapping of clusters in local minima, and uses a density grid in place of pairwise repulsive forces to speed up execution. Computation times are thus order O(N) rather than $O(N^2)$. Another advantage of the VxOrd algorithm is that it determines the number and size of clusters automatically based on the data input. Plus, rather than placing objects in discrete (round) clusters, VxOrd often gives elongated or continuous structures (which look like ridges in a landscape visualization) that bridge multiple fields. The VxOrd FDP does not accommodate a continuous stream of updated data, as do some other FDP's.

Semantic Treemaps, recently proposed by (Feng & Börner, 2002), are another option to apply FDP to handle large data sets. Semantic tree maps apply clustering techniques to organize documents into clusters of semantically similar documents. Subsequently, the tree map approach (Shneiderman, 1992) is utilized to determine the size (dependent on the number of documents) and layout of clusters. Finally, FDP is applied to the documents in each cluster to place them based on their semantic similarity. By breaking the data set into smaller chunks, the computational complexity of FDP is reduced at the cost of global optimality.

HyperSpace, formerly Narcissus, used FDP to visualize hyperlinks among Web pages (Hendley, Drew, Wood, & Beale, 1995). FDP has been used on small data sets by Börner (Katy Börner, 2000; Börner et al., 2000). Much larger literature (Boyack et al., 2002), patent (Boyack, Wylie,

Davidson, & Johnson, 2000), and even genomic (Kim et al., 2001) data sets have been clustered using the VxOrd FDP.

4.4 Visualization and Interaction Design

Given data objects and their spatial positions, visualizations need to be designed that can be intuitively understood and effectively and accurately explored by a human user. However, nobody should expect to understand a complex visualization in a few seconds. "The first response should be content related, not layout." (Eduard Tufte, 1998).

Different frameworks and taxonomies to characterize information visualization techniques have been proposed. Most commonly used is Ben Shneiderman's 1996 framework characterizing IV in terms of data types and user tasks to "sort out the prototypes [that currently exist] and guide researchers to new opportunities.". The framework defines:

- ***Data types*** comprising linear, planar, volumetric, temporal, multidimensional, tree, network, and workspace.[24]
- ***Typology of Tasks*** such as overview, zoom, filter, details-on-demand, relate, history, and extract.
- ***Visualizations*** resemble landscapes, circle plots, term plots, spotfires, starfields, etc.
- ***Necessary features*** comprise interaction, navigation, detail on demand, etc.

Subsequently, we review visualization as well as interaction design techniques and approaches.

4.4.1 Visualization

Visualization refers to the design of the visual appearance of data objects and their relationships. Well-designed domain visualizations:

- Provide an ability to comprehend huge amounts of data on a large-scale as well as a small-scale.
- Reduce visual search time (e.g., by exploiting low level visual perception).
- Provide a better understanding of a complex data set (e.g., by exploiting data landscape metaphors).
- Reveal relations otherwise not noticed (e.g., by exploiting perception of emergent properties).
- Enable a data set to be seen from several perspectives simultaneously.
- Facilitate hypothesis formulation.
- Are effective sources of communication.

Information visualization—the process of analyzing and transforming non-spatial data into an effective visual form—is believed to improve our interaction with large volumes of data (Card et al., 1999; Chen, 1999a; Gershon, Eick, & Card, 1998; Spence, 2000, 2001). One major key element of any successful visualization is to exploit visual perception principles. Books by Ware (2000) and Palmer (1999) provide excellent introductions to the subject. Visualizations help an increasingly diverse and potentially non-technical community to gain overviews about general patterns and trends and to discover hidden [semantic] structures. In addition, complex

[24] Added in his textbook (Shneiderman, 1997).

visualizations of different viewpoints of thousands of data objects can greatly benefit from storytelling (Gershon & Ward, 2001). Storytelling and sharing is a powerful human strategy to teach effectively, to stimulate critical and creative thinking, and to increase awareness and understanding. Last but not least, the design and presentation of meaningful visualization are an art that requires years of expertise and diverse skills. However, the visual perception, story telling, and artistic aspects of visualization design are beyond the scope of this paper.

4.4.2 Interaction Design

Interaction design refers to the implementation of techniques such as filtering, panning, zooming, distortion, etc. to efficiently search and browse large information spaces.

Ben Shneiderman at the University of Maryland proposed a mantra to characterize how users interact with the visualization of a large amount of information: *Overview, Zoom-in (Filter), and Details on Demand* (Shneiderman, 1996). Users would start from an overview of the information space and zoom in to the part that seems to be of interest, call for more details, and so on. The term "drill down" is also used to refer to processes equivalent to the "zoom in" part of the mantra. As for where to zoom in, theories such as optimal information foraging (Pirolli & Card, 1999) appear to be a promising route to pursue.

To issue meaningful queries or to exploit labeling of maps, users need a working knowledge of the subject domain vocabulary. Given the imprecise nature of human language, users frequently encounter the "vocabulary mismatch problem" (Chen et al., 1998; Deerwester et al., 1990).

Although domain maps might provide searching facilities – e.g., documents matching a query are highlighted – one of their main purposes is to support browsing – i.e., the exploration of an information space in order to become familiar with it and to locate information of interest. "Browsing explores both the organization or structure of the information space and its content." (Chen et al., 1998) It requires working knowledge of the applied knowledge organization (typically alphabetical, categorical, or hierarchical) and how to navigate in it. To ease navigation, numerous (real-world) visualization metaphors have been proposed and applied to help improve the understanding of abstract data spaces. Among them are 2-D "cartographic maps," 2-D/3-D "category maps," "desktop," and 3-D "landscape" or "star field" visualizations.

Paul Dourish and Matthew Chalmers identified three major navigation paradigms: spatial navigation – mimicking our experiences in the physical world; semantic navigation – driven by semantic relationships or underlying logic; and social navigation – taking advantage of the behavior of like-minded people (Dourish & Chalmers, 1994). Ideally information visualization facilitates and supports all three.

4.4.3 Focus+Context

The desire to examine large information spaces on small displays with limited resolution leads to the development of different focus and context techniques that enable users to examine local details without losing the global structure. Distortion-based techniques keep a steady overview. They enlarge some objects while simultaneously shrinking others. Ideally, the total amount of information displayed can be set flexibly and is constant even when users change their focuses of attention over several magnitudes.

Hyperbolic Trees, developed at Xerox PARC, were one of the very first focus and context techniques (Lamping, Rao, & Pirolli, 1995). Based on Poincare's model of the (hyperbolic) non-Euclidean plane, the technique assigns more display space to a portion of the hierarchy while still embedding it in the context of the entire hierarchy. A 3-D hyperbolic viewer was developed by Tamara Munzner (1997; 1998).

Hyperbolic trees are very valuable to visualize hierarchical structures such as file directories, Web sites, classification hierarchies, organization hierarchies, newsgroup structures, etc. While traditional methods such as paging (dividing data into several pages and displaying one page at a time), zooming, or panning show only part of the information at certain granularity, hyperbolic trees show detail and context at once. Although hyperbolic trees have not been used to visualize knowledge domains, they are commonly used with patent trees, and might be well suited to visualization of other hierarchical data.

Fisheye views developed by George Furnas (1986) show a distorted view of a data set in an attempt to show local detail while maintaining global context. They mimic the effect of a wide-range fisheye camera that shows the whole world, but have higher magnification in the focus center and shrink objects in relation to their distance to the center of focus.

Two transformation options can be applied to the fisheye view: Cartesian and polar. For Cartesian transformation, all the regions are rectangular. Polar transformation regions can be arbitrarily shaped. The technique was improved by Sarkar and Brown (1994) with respect to layout considerations. Fisheye views have also been applied to improve ET-maps (Yang et al., 1999).

Fractal views, based on Mandelbrot's fractal theory (1988), were first applied to the design of information displays by Hideki Koike (1993). They can be utilized to abstract displayed objects and to control the amount of displayed information based on semantic relevance by removing less important information automatically[25]. The fractal dimension, a measure of complexity, is used to control the total number of displayed nodes.

Fractal Views have been applied to visualize huge hierarchies (Koike & Yoshihara, 1993) and to control the amount of information displayed in ET-Maps (Yang et al., 1999).

Semantic zoom was also introduced by Furnas, and provides multiple levels of resolution. The view changes depending on the "distance" the viewer is from the objects. Semantic zoom was implemented in MuSE – Multiscale editor (Furnas & Zhang, 1998). It was also used in the Galaxy of News system that visualizes large quantities of independently authored pieces of information such as news stories (Rennison, 1994).

Zoomable user interfaces, also called ZUIs, incorporate zooming as a fundamental aspect of their design. They place documents at absolute positions within a large zoomable space.

[25] A linear clustering algorithm that groups objects according to the effects they have on the fractal dimension of the clusters was proposed in (Barbará & Chen, 2000).

Combined with animated navigation, this helps to give users a sense of structure and of where they are within a large information space.

Pad++ is an environment for exploring zooming techniques for interfaces (Bederson et al., 1996). Jazz is a Java 2 toolkit (http://www.cs.umd.edu/hcil/jazz/) developed by Ben Bederson and his colleagues at the University of Maryland. It supports the development of 2D structured graphics programs in general, and ZUIs in particular.

4.5 Discussion

Table 1 provides an overview of main features of the dimensionality reduction and ordination techniques of section 4. Among the features are scalability, computational cost[26], interpretability of dimensions, dynamic or static layout, and the scale (global or local) to which it can optimally be applied.

Table 1: Comparison of Techniques

Technique	Scalability	Computation costs	Interpret. Dim	Layout	Optimality scale
Eigenvalue	high	high	often	static	global
FA/PCA	limited	medium	often	static	global
MDS	limited	medium	often	static	global
LSA	high	high	no		global
PFNet	medium	medium	no	static	Local or global - depends on parameter setting
SOM	high	high	no	static	global
Triangulation	medium	medium	--	static	local
FDP	limited	high	--	dynamic	local

One exception to Table 1 is noted here. Although VxOrd is classified as a FDP algorithm, it does not act like FDPs as characterized in the table. Rather, it can scale to very large data sets (millions of similarity pairs, only limited by memory constraints), has very fast run times, and provides a static layout. Some ordination techniques have a very high computational complexity for large data sets. This complexity can be reduced at the cost of global optimality by breaking the data set into smaller chunks, ordinating each cluster, and compiling all clusters into a single map.

Incrementally updated visualizations of domains based on incremental data updates, or perhaps continually updated visualizations based on sequential streaming of previously extracted data are very desirable for domain analysis see section 7 on Promising Avenues of Research. However, few current techniques (only some FDPs) can accommodate this type of data.

[26] Most data analysis techniques are computationally expensive and are applied in a batch job. During run time the results of the data-mining step are used to interactively visualize a data set under a certain point of view.

5 THE ARIST DATA SET

5.1 Data Retrieval

To demonstrate the literature mapping process and to show the different measurements, layout routines, and visualization metaphors that may be used to visualize a knowledge domain in action, we have developed a data set – named ARIST data set - consistent with the subject of this chapter.

First, data were retrieved from the Science Citation Index (SCI) and Social Science Citation Index (SSCI) by querying the titles, abstracts, and terms (ISI keywords and keywords plus) fields for the years 1977-July 27, 2001. Query terms and the number of records retrieved for each query are shown in Table 2.

Table 2. Search terms used to generate the ARIST bibliographic data set.

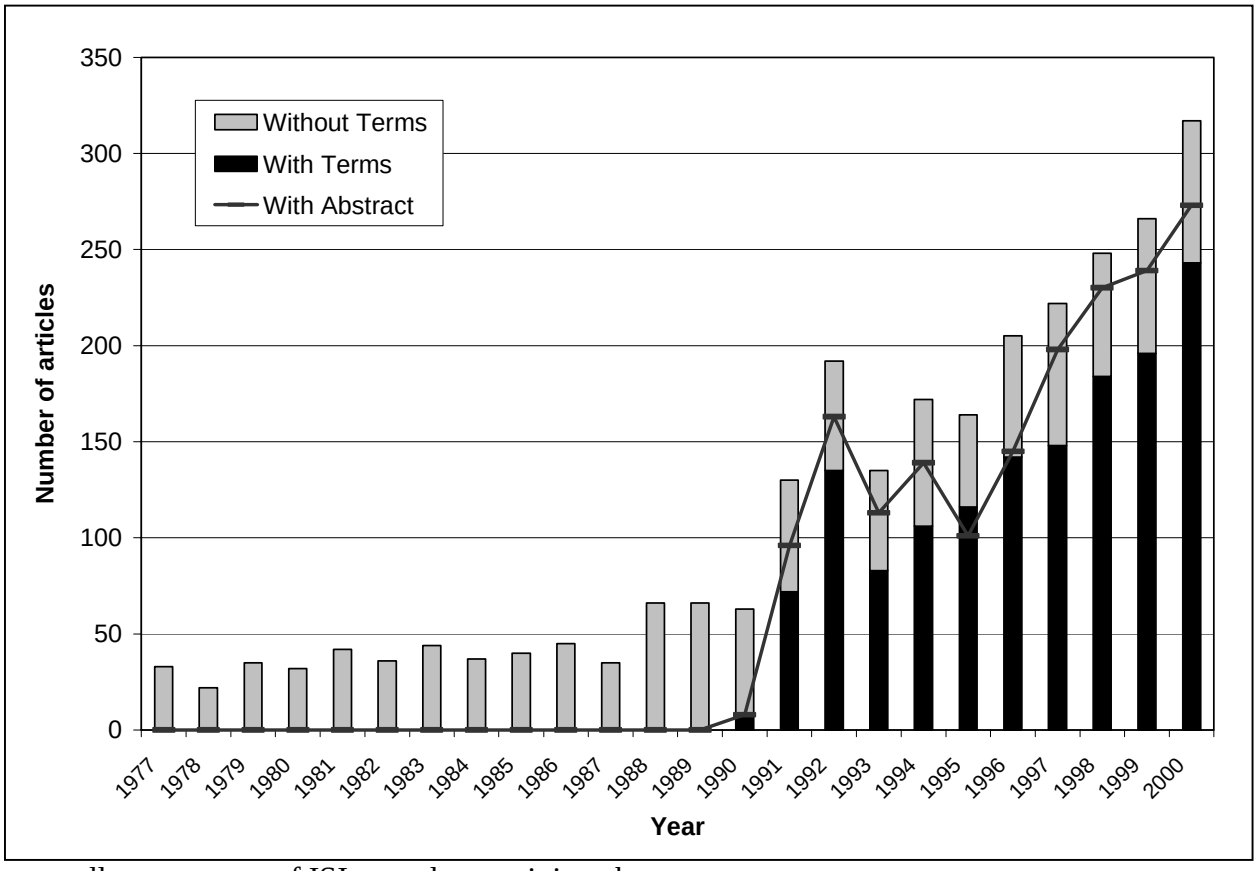
SEARCH TERM	Number
Topic Citation Analysis:	
citation analysis	596
cocitation OR co-citation	177
co-occurrence AND (term OR word)	77
co-term OR co-word	52
science map[ping] OR mapping science OR map[ping] of science	32
Topic Semantics:	
semantic analysis OR semantic index OR semantic map	331
Topic Bibliometrics:	
bibliometric	818
scientometric	327
Topic Visualization:	
data visualization OR visualization of data	275
information visualization OR visualization of information	113
scientific visualization	268

Search terms included terms relevant to *citation analysis*, *semantics*, *bibliometrics*, and *visualization* to allow overlaps between those terms and fields to be shown. These four fields will be referred to extensively in the domain analyses of section 6. Of the 2764 unique articles retrieved, 287 were retrieved by more than one of the query terms.

5.2 Coverage

It is extremely important to choose an appropriate data source for retrieval, one whose data are likely to provide answers to the questions one wishes to answer using domain visualization. As an example, we discuss some limitations associated with the data for our ARIST data set. Numbers of articles retrieved by year are shown in Figure 2 for two categories: articles with terms (ISI keywords) and articles without terms. As is well known, ISI's databases did not include either abstracts or terms prior to 1991. Thus, any maps based on either abstract text or terms will naturally exclude any articles prior to 1991. In addition, as shown in Figure 2, terms are available for only 71% of the articles published since 1991. (The lack of terms can be due to

several things; for instance, the journal may not require index terms from the authors, the journal may not supply terms to the database vendor, or the database vendor may choose not to index an article.) This makes the use of terms a less-than-optimum basis for mapping of these data. By contrast, abstracts are available for 84% of the post-1991 ARIST data set, making it a richer source of information. The percentages listed here apply only to this data set; we do not know the overall percentages of ISI records containing abstracts or terms.



Book, journal and/or conference coverage can also be an issue. For instance, *JASIS(T)*, *Scientometrics*, *Journal of Information Science*, *Information Processing and Management*, and the *Journal of Documentation* are key sources for visualization of science or knowledge domains (see Table 3). Yet the SCI only started coverage of these journals in the mid-1990s. (*JASIS*, *Scientometrics*, and *JDoc* were covered through the mid-1980s.) Queries to the SCI alone would not have provided sufficient coverage of the intended fields over this time period due to this lack of key journal coverage. Thus, we queried the SSCI as well, which covered those journals over the entire time period under investigation. Users of ISI's Web of Science can conduct a search across all databases such as the SCI, SSCI, and the Arts and Humanities Citation Index (AHCI).

Figure 2. Numbers of articles in the ARIST data set by year with terms (ISI keywords) or abstracts.

Table 3. Number of articles by journal in the ARIST set (10 or more articles per journal).

Journal	Categories	# Papers
Scientometrics	LIS, CS	482
JASIS(T)	LIS, CS	139
Journal of Information Science	LIS, CS	51
Information Processing & Management	LIS, CS	45
Lecture Notes in Computer Science	CS	39
Research Policy	Other	32
Journal of Documentation	LIS, CS	31
Current Contents	Other	30
Computers & Graphics	CS	27
IEEE Transactions on Visualization and Computer Graphics	CS	25
Bulletin of the Medical Library Association	LIS	25
IEEE Computer Graphics and Applications	CS	20
Medicina Clinica	Other	20
Library & Information Science Research	LIS	19
Social Studies of Science	Other	18
Computer	CS	16
Computer Graphics Forum	CS	16
Libri	LIS	16
Lecture Notes in Artificial Intelligence	CS	15
Future Generation Computer Systems	CS	15
International Forum on Information and Documentation	LIS	15
Landscape and Urban Planning	Other	14
Proceedings of the American Society For Information Science	LIS	14
Proceedings of the ASIS Annual Meeting	LIS, CS	14
Nachrichten Fur Dokumentation	LIS	14
Library Trends	LIS	13
Library Quarterly	LIS	12
Science Technology & Human Values	Other	12
Scientist	LIS	12
Library and Information Science	LIS	12
Omega-International Journal of Management Science	Other	11
Computers & Geosciences	CS	10
Zentralblatt Fur Bibliothekswesen	LIS	10

Table 3 also shows that the data were dominated by journals jointly classified in the Library & Information Science (LIS) and Computer Science (CS) categories. Journals classified as "Other" in Table 3 come from a variety of categories, and suggest that the fields covered by our original queries are accessed by many other disciplines.

For completeness, we include a distribution of the number of articles per field per year, along with average citation counts, for the ARIST data set (see Figure 3). This shows the dramatic increase in publishing in citation analysis and bibliometrics starting in the late 1980s, and the birth of the visualization field around the same time. It also shows that citation analysis articles were more highly cited that bibliometrics articles in the 1970s and 1980s, and that citation counts for all four fields have generally dropped throughout the 1990s. The most recent articles have, of course, been cited infrequently due to their young age.

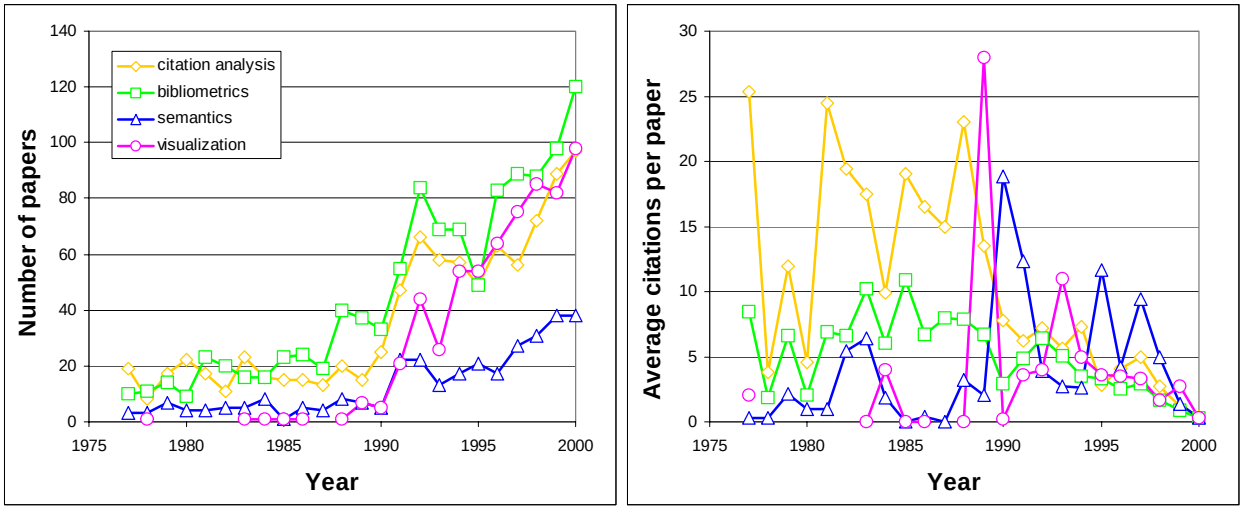


Figure 3. Numbers of articles by field per year in the ARIST data set with average citation counts. Articles contribute to counts in more than one field if retrieved by queries from multiple fields.

Data coverage issues also raise other questions such as:

- Does lack of appropriate coverage cause significant distortions of domain visualizations?
- If so, to what extent are the levels of quality of the final domain visualization and the analysis results undermined?
- Are there ways to compensate for missing data?

We have no ready answers to these questions, but suggest that they are important topics for discussion and further research.

6 THE STRUCTURE OF THE SUBJECT DOMAIN

As described in section 4, there are many mapping techniques available to work side by side on the same data and produce images of a domain from different perspectives. This allows us to stitch different pictures into a bigger one, which will reveal more insights about a domain than use of just a single technique. Multiple tools enhance the utility of domain visualization.

6.1 Multiple Maps of the Domain

The overall organization of the field of domain visualization (the overall subject of the ARIST data set) takes advantages of several emerging techniques. These techniques make it possible not only to decompose the domain according to a range of quantitative measures, but also to compare and contrast different pictures of the same domain.

For example, in GSA and Starwalker, the use of factor analysis allows us to break down a domain network into components. A long recognized advantage of using factor analysis over traditional clustering techniques is that factor analysis does not force us to classify one object

into one cluster or the other; instead, it preserves an important type of phenomenon: truly important work is often universal. In SOM, the overall structure is depicted in forms of adjacent regions. Therefore, matches and mismatches between various versions of the domain maps will provide insights. VxInsight uses a landscape metaphor and portrays the structure of a literature space as mountain ridges of document clusters. The size of a cluster and its relative position in the layout provide valuable clues to the role of the cluster in the overall structure. Many different snapshots of the ARIST domain are explained in the remainder of this section.

6.1.1 ARIST-GSA/StarWalker

We generated an author co-citation analysis (ACA) map and a document co-citation analysis (DCA) map based on the ARIST data set using the four-step procedure described in (Chen & Paul, 2001).

Data Preparation

First, we selected authors whose work has received citations above a determined threshold for the author co-citation visualization. This selection was on the first-author only basis due to the availability of the information.[27] Documents were selected similarly for the document co-citation visualization. The threshold parameter can be increased or decreased to control the number of authors or documents to be analyzed. Conventionally, the author citation threshold is set to 10. The intellectual groupings of these authors provide snapshots of the underlying knowledge domain. We computed the co-citation frequencies for these authors from a citation database, such as ISI's SCI or SSCI. ACA uses a matrix of co-citation frequencies to compute a correlation matrix of Pearson correlation coefficients. Some researchers believe that such correlation coefficients best capture an author's citation profile.

Author Co-citation Analysis

Second, we applied Pathfinder network scaling to the network that the correlation matrix defines. Although factor analysis is a standard ACA practice in traditional author co-citation analysis, MDS and factor analysis rarely appear in the same graphical representations. We then overlay the intellectual groupings that factor analysis identifies and the interconnectivity structure of a Pathfinder network. Authors with similar colors essentially belong to the same specialty and should appear as a closely connected group in the network. Therefore, we can expect to see the two perspectives converge in the visualization.

Finally, we display the citation impact of each author atop the intellectual groupings. The height of a citation bar—which consists of a stack of color-coded annual citation sections—represents the magnitude of the impact. Sample author co-citation analysis maps are displayed in Figure 4 and Figure 5.

The factor analysis identified 10 factors whose eigenvalues are greater than one. These factors explain 90% of variance in the data. Each factor corresponds to a specialty in domain

[27] This way of gathering data has the effect of privileging sole or first authors. Researchers who publish frequently in non-first author positions are thus not included even though they should be. Research has shown that first-author citation studies distort the picture in terms of most influential researchers. All-author citation counts should be preferred when visualizing the structure of research fields. The subfield structure tends to be just about the same for both methods (Persson, 2001; van Dalen & Henkens, 2001).

visualization. The largest three factors cumulatively explain 63% of variance. The following four specialties can be identified in the map, although the Pathfinder structure and the overlay factor analysis color scheme did not converge in this case – the sign of a heterogeneous subject domain:

- mapping science: fundamentals,
- social studies of science,
- bibliometrics: quantitative analysis and evaluation, and
- scholarly communication and co-citation analysis.

The top-three specialties correspond to color-coded factors in the map: mapping science in red, social studies of science in green, and bibliometrics in blue. Remaining specialties are likely to be a combination of all the colors, and readers can cross-reference between factor analysis results and the map.

The resultant author co-citation map contains 380 authors who have nine or more citations over the entire period between 1977 and 2001. Pathfinder network scaling limited the number of "salient" connections among these authors to 384. As usual, each author's node is colored by the factor loadings in the largest three specialties. An author co-citation map of a focused, coherent subject domain should demonstrate a considerable degree of conformance between the Pathfinder network structure and the factor analysis color patterns. However, this is not the case here, suggesting that science mapping constitutes a number of largely independent disciplines - as (Leydesdorff & Wouters, 2000) describes the status of scientometrics - this is pre-paradigmatic.

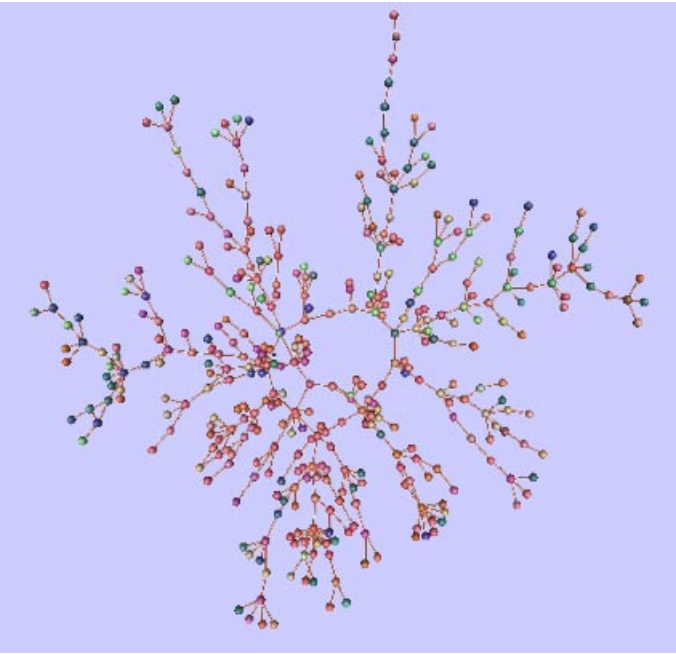


Figure 4. An overview of the author co-citation map (1977-2001), consisting of 380 authors with 9 or more citations. The map is dominated by the largest specialty of citation indexing. No strong concentration of other specialties are found, which implies the diversity of the domain.

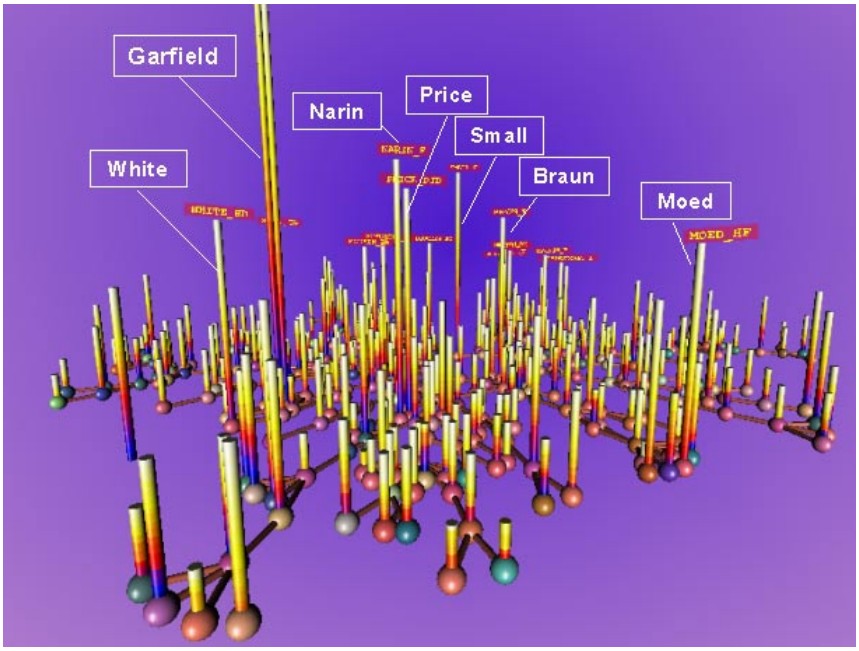


Figure 5. A landscape view of the ACA map displayed in Figure 4. The height of a citation bar indicates the number of citations for the correspondent author. The spectrum of colors on each citation shows the time when citations were made. Authors with more than 50 citations are displayed with semi-transparent labels.

It is interesting to note that none of the four largest factors are centered in either visualization or semantics, which were two of the four groups comprising the ARIST data set. These two fields are relatively new and are not highly cited (see Figure 3). Thus, they are unlikely to be strong factors in an ACA-type analysis.

Document Co-citation Analysis

Given the flexibility of GSA, we also generated a document co-citation map based on the top sliced set of documents (see Figure 6 and Figure 7). As expected, there are as many as 15 factors of which eigenvalues are greater than one. These 15 factors cumulatively explained 90% of variance in the ARIST data. The largest four factors explained 56%. The following four specialties are relatively easy to identify in the map, corresponding to the largest four factors respectively:

- mapping science: fundamentals,
- social studies of science,
- bibliometrics: quantitative analysis and evaluation, and
- scholarly communication and co-citation analysis.

As with the author co-citation maps, the top-three specialties in document co-citation maps also correspond to color-coded factors in the map: mapping science in red, social studies of science in green, and bibliometrics in blue. Remaining specialties are likely to be a combination of all the colors, and readers can cross-reference between factor analysis results and the map.

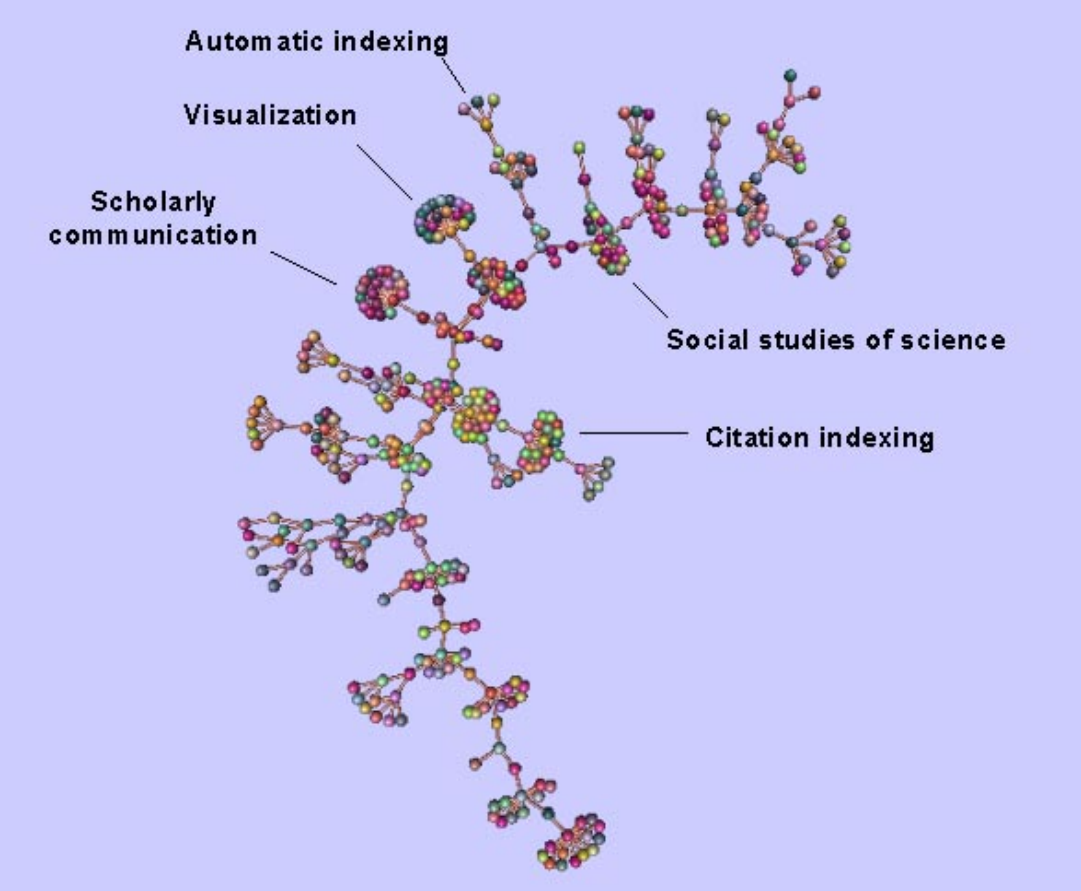


Figure 6. An overview of the document co-citation map of 394 articles with 10 or more citations. In this network, a number of tight clusters of documents are connected to an artery-like chain. Documents on the artery chain tend to be seminal works of connected clusters. For example, Diana Crane's Invisible College (Crane, 1972) connects the scholar communication cluster to the artery chain.

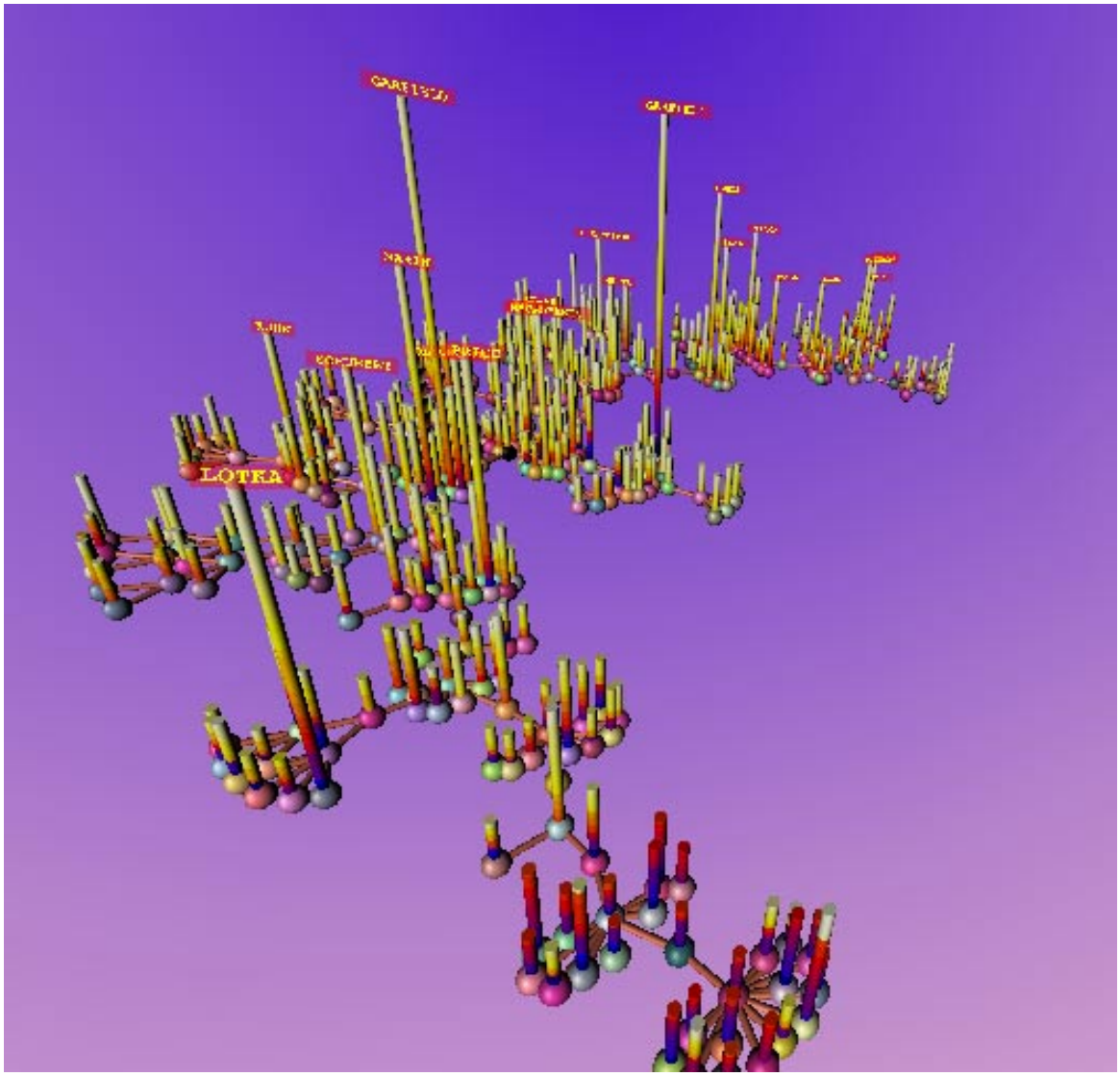


Figure 7. A landscape view of the DCA map displayed in Figure 6 at distance.

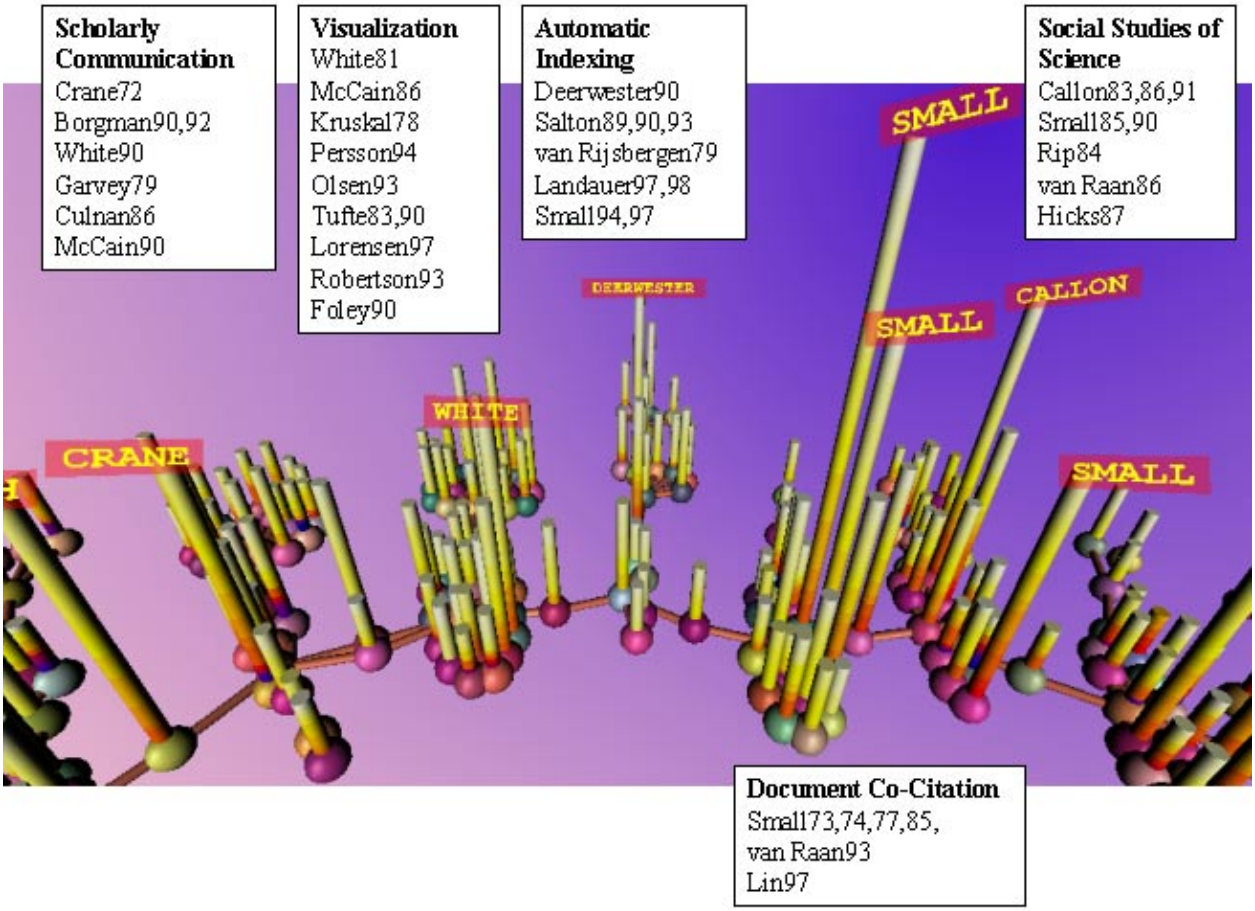


Figure 8. A close-up view of a few clusters along the main artery in the DCA map. The height of a bar represents the number of citations to a publication. Labels indicate articles in clusters, for example, Small73 for an article of Small in 1973. Multiple publications within the same year are not distinguished at this level. For example, Small73 includes all Small's publications in 1973.

6.1.2 ARIST-ET-Map

Bin Zhu and Hsinchun Chen of the University of Arizona visualized the ARIST data set using ET-Maps (see section 4.3.2). They trained 10*10 nodes using the ID/keyword data of the ARIST data set. At the end of the map creation process, each node is associated with a list of documents that are semantically similar to each other. In addition, a phrase was assigned to each node as its label and adjacent nodes that have the same label are grouped into one region or category. Thus, spatial proximity on the map indicates semantic proximity, meaning if two categories are close to each other on SOM map, they are also semantically similar to each other.

A screenshot of the resulting ET-Map utilizing different visual encasings is shown in Figure 9. The top-level map shows 14 subject regions represented by regularly shaped tiles. Each tile is a visual summary of a group of documents with similar keywords. The tiles are shaded in different colors to differentiate them, while labels identify the subject of the tile. The subjects are also

listed on the left hand side together with a number in brackets telling how many individual documents it contains. In a typical browsing session a user would get an overview first, zoom into areas of interest, and access a particular document. Alternatively, a user can select a category of interest and the interface will display documents within that category.

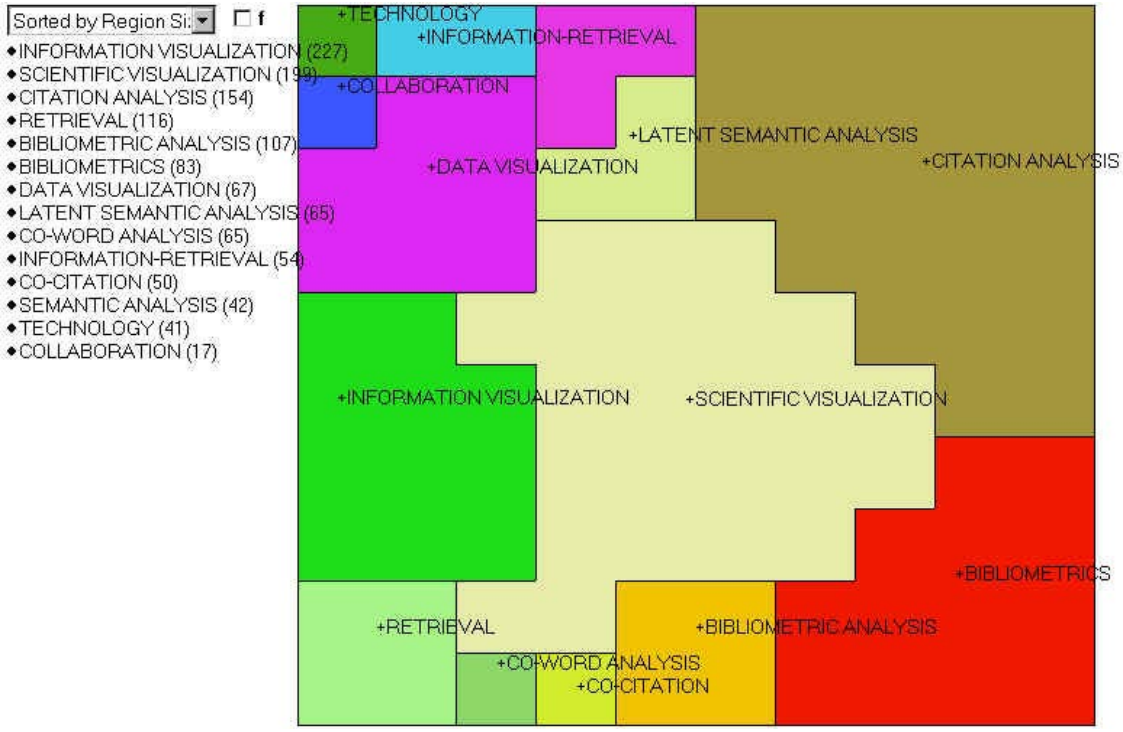


Figure 9. ET-Map of the ARIST data set using keywords. Left is a list of labels of all categories. A user can select a category of interest and the interface will display documents within that category.

Note that subject area *citation analysis* appears to be much larger than *information visualization* even though the citation analysis area has fewer documents. The size of the subject area is not necessarily related to the number of documents in an ET-map, but rather it denotes the amount of space between areas based on the number of nodes used to generate the map (see Figure 14b in section 6.2).

6.1.3 ARIST-Cartographic-SOM Maps

André Skupin, a geographer at the University of New Orleans, uses SOMs to generate domain visualizations in a cartographic fashion (Skupin, 2000; Skupin & Buttenfield, 1996). He used SOM_PAK to train a SOM (see section 4.3.2) based on the ID/keyword list of the ARIST data set and ArcGIS to generate the visualization. Labeling of clusters was done automatically within ArcGIS based on rules that were given to it regarding the link between attributes and label characteristics.

The map shown in Figure 10 aims at facilitating the use of the same skills traditionally associated with geographic maps. The underlying SOM consists of 2200 neurons (i.e. 40x55 nodes).[28] A hierarchical clustering tree was computed for these neurons in order to allow scale-dependent exploration of the data set in a zoomable interface. In this particular visualization, only the 25-cluster solution is shown.

For each cluster, three label terms are computed, to better convey cluster content. Since the clusters were computed from the SOM itself, these labels indicate the potential topic or theme that one would expect to find with articles assigned to a particular cluster. In order to show the relative prominence of topics in the data set, the clusters were ranked according to the number of articles of this data set that were actually assigned to them. Clusters containing more articles appear larger and more prominent.

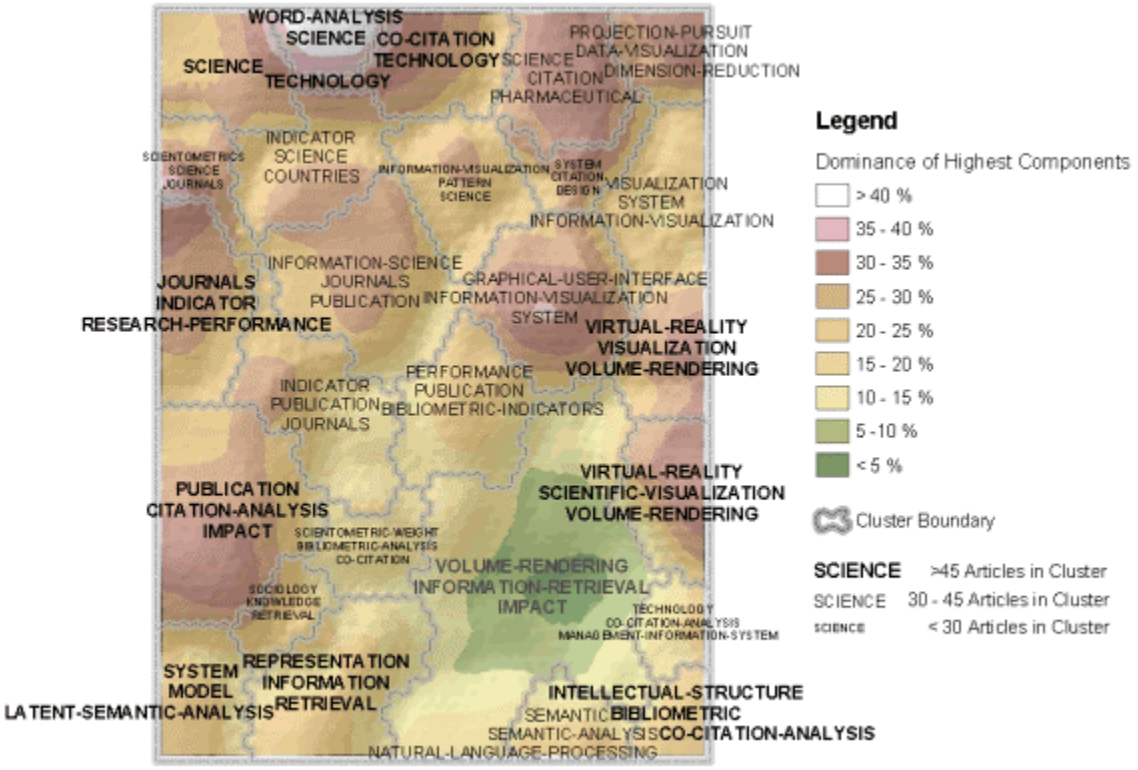


Figure 10. Cartographic SOM map of ARIST data set.

The terrain visualization expresses the degree to which the three highest-ranked components of each neuron dominate the neurons n-dimensional term vector. The purpose of this is to allow some judgment regarding the relative merits of the clustering solution that is overlaid. Higher elevation—i.e., percentage—indicates a very organized, focused, and coherent portion of the information space. These areas tend to be recognized and preserved by the cluster routine, which shows up nicely in this map. On the other extreme, there is the low-lying area, especially at the bottom right of the map. This indicates a lack of strong organization, a lack of distinct themes that would be recognized by the clustering routine. What is important here is that this does not

[28] Thus, SOMs use a raster data model instead of a vector data model.

mean that there is necessarily a lack of "meaning" to this area. Rather, clusters containing it should be interpreted more cautiously. In this case, the main reason for the lack of organization is that most of the articles with 1-2 keywords are congregated here. In the 25-cluster solution this area is contained in a cluster whose labels ("Volume-Rendering" – "Information-Retrieval" – "Impact") indicate a lack of coherence. Labels for this cluster are scaled correctly, but de-emphasized by using a gray font color. The clustering solution and terrain visualization also indicate areas of transition and overlaps between different major topics. For example, note the cluster labeled "Information-Visualization" – "Pattern" – "Science" that is located between the larger clusters that are dominated by "Information Visualization" and "Science."

Various ways in which cartographic techniques could be used to improve information visualization are discussed in (Skupin, 2000) see also section 7.

6.1.4 ARIST-VxInsight

Sandia's VxInsight was used to generate a number of document-based views of the ARIST data set using several different similarity functions. Four separate maps were created, all using the VxOrd FDP algorithm:

1. A citation-based map using direct and co-citation linkages after the combined linkage method of Small (1997) using a direct:cocitation weighting factor of 20:1. Four different time segments are shown in Figure 11.
2. A co-term map based on a cosine similarity using the ISI keywords (see Figure 12, left).
3. A map based on LSA over words extracted from the titles of articles. LSA was performed using SVDPACK (see section 4.1.2) to generate a document-by-document similarity matrix. Only similarity values ≥ 0.9 were used in the VxOrd FDP to generate the map shown in Figure 12, right.
4. A co-classification map based a cosine similarity from the ISI journal classifications for each article (see Figure 13).

In all three figures we used the following color scheme to indicate the various query terms that have been used to retrieve articles: white for citation analysis, co-citation, co-word, etc.; green for bibliometrics and scientometrics; blue for semantics; and magenta for visualization. Articles that match multiple query terms show multiple color markers.

A quick perusal of the landscapes in Figure 11-Figure 13 shows that each reveals different information about the domain. The citation-based map in Figure 11 shows the relationship between the four fields for four different time segments. It is very easy to see the growth in the four fields – citation analysis, bibliometrics, semantics, and visualization – from a comparison of the figures. Citation analysis and bibliometrics (white and green dots) both have roughly equal numbers of papers during the first time segment (1977-82). More detailed analysis reveals that citation analysis was stronger in the late 1970s with bibliometrics picking up steam in the early 1980s.

The second segment (1983-88) shows bibliometrics as the larger field with some well defined clusters near the top of the map (labeled A). The semantics and visualization fields have not yet appeared (see Figure 3 as well). The third segment (1989-94) shows the formation of the

visualization field in three clusters (labeled B). The fact that these visualization clusters are formed at the edges of a citation map indicates that they are not well linked to the main body of citation and bibliometric work in the center of the map. Pictures showing the citation links (not shown here) confirm this and the fact that the visualization papers are not as highly cited as the citation and bibliometrics papers. This may be due to their relatively young age or perhaps may indicate that visualization researchers tend to cite other work less frequently.

The fourth segment (1995-2000) shows semantic analysis as a new field (labeled C), which, like the visualization clusters, is not well linked to the main body of work. The citation analysis and bibliometrics fields each seem to be well defined, with some mixing of the two in certain clusters. While bibliometrics seemed to be growing faster than citation analysis in the middle two segments (late 80s and early 90s), comparison of the numbers of papers in the late 90s shows citation analysis to be gaining strength. There is also an interesting cluster (labeled D) in which citation analysis, bibliometrics, and visualization are mixed. Additional analysis on this and surrounding clusters will be given in section 6.2 (see Figure 14).

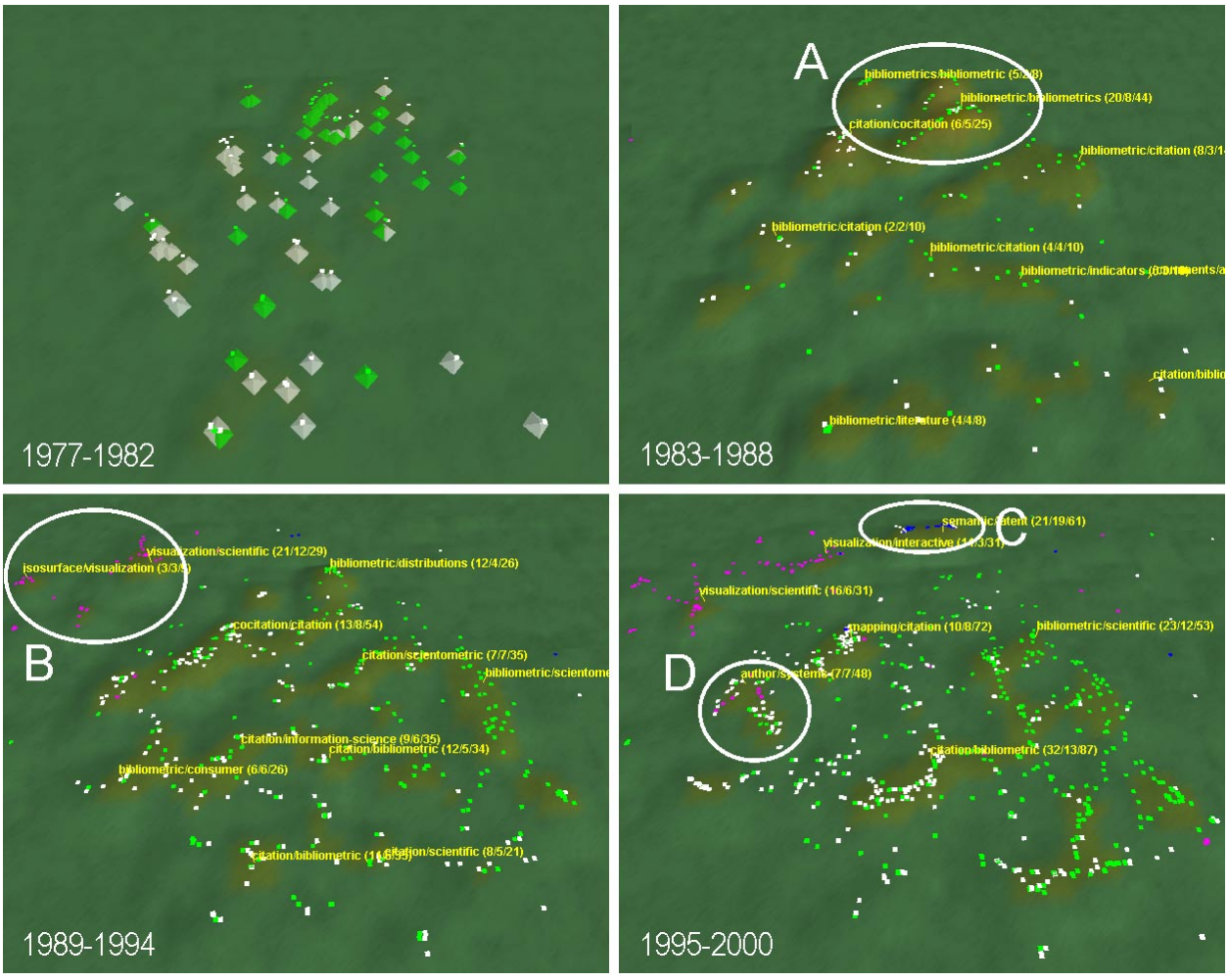


Figure 11. VxInsight citation maps of ARIST data for four different time segments. Circles indicate areas highlighted in the text. Dot color legend – WHITE: citation analysis, GREEN: bibliometrics, BLUE: semantics, MAGENTA: visualization.

Figure 12 shows both the co-term and LSA maps. In the co-term map, citation analysis and bibliometrics articles are found mainly in the two large peaks at the right edge of the map, and seem to be more mixed than in the citation map of Figure 11. This could be simply because there are fewer clusters in the co-term map than in the citation map. However, it could also indicate a lack of specificity in using a term-based similarity, which leads to fewer clusters. The visualization and semantics papers form clusters of their own. There are a couple of peaks where the four fields are mingled in the center of the landscape.

The map based on LSA over article titles shows some different features. The titles contained 4,802 unique words, while abstracts contained 14,494 unique words. Here, the semantics papers are spread throughout the landscape. The visualization papers are also more mixed, and in several cases, they seem to bridge groups of citation analysis and bibliometrics articles. Bibliometrics and citation analysis articles, while appearing in the same clusters, are segregated within the clusters.

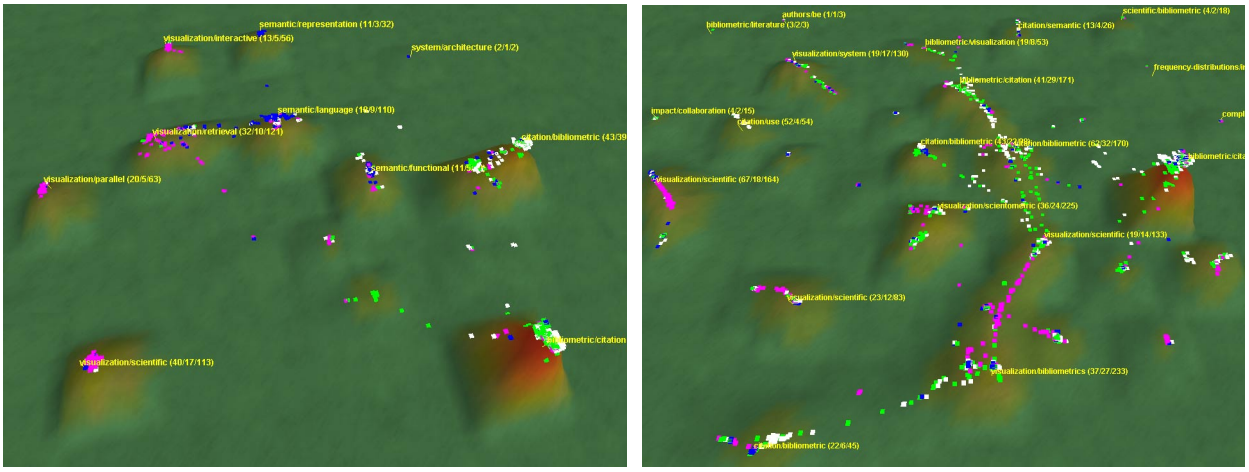


Figure 12. VxInsight co-term (left) and LSA (right) maps of ARIST data

The co-classification map in Figure 13 gives completely different information about the domain. This type of map has use in that one can clearly see the fields in which various groups of articles were published. The large peak with white and green dots (labeled A) is dominated by information and library sciences, which is where one would expect citation analysis and bibliometrics articles to be published. However, the white/green combination is also found in peaks comprised mainly of planning and development journals (labeled B), and general medicine (labeled C). Interestingly, a peak comprised mainly of articles in neuroscience-related journals (labeled D) contains articles from all four main areas of the ARIST data.

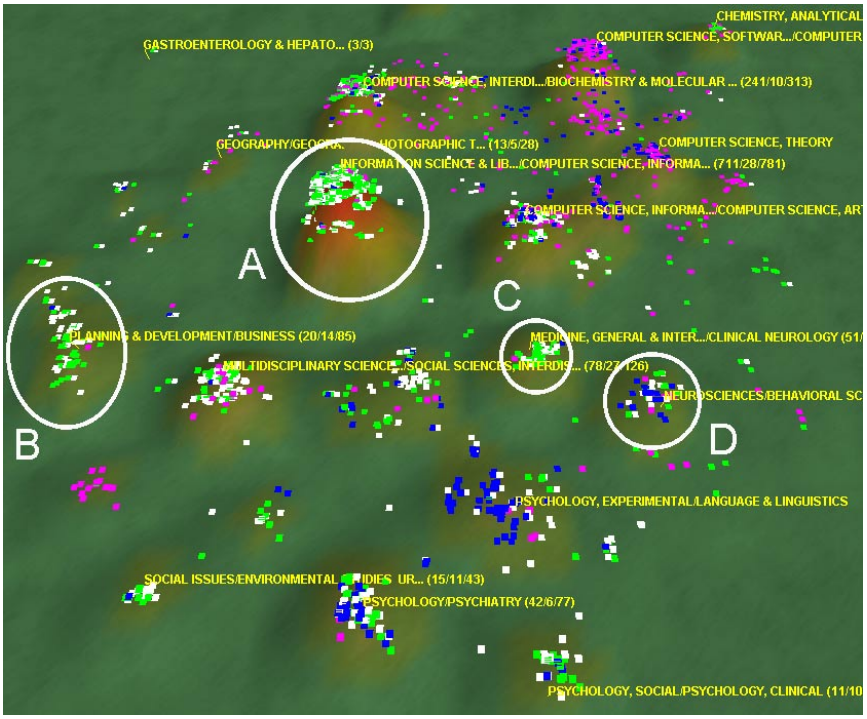


Figure 13. VxInsight co-classification map of ARIST data

Note that the VxInsight tool is not just a way to take pictures of a domain visualization, but is a dynamic tool with which one can browse the information. Zoom, rotation, dynamic labeling (updated at each zoom step) based on different attributes such as titles, terms, etc., showing of citation or other linkages, filtering by date, detail on demand for individual articles, the ability to import different layouts, and a query facility are all features that enable a highly interactive exploration of the information space.

6.2 Comparison of Maps

The layouts of different [document] maps – Cartographic SOM map, ET map, VxInsight co-term map and LSA map – based on terms or words are compared here. Three of the maps were generated from an ID/term list, as shown in Table 4, and one was generated from LSA on article titles using VxOrd. Information about the citation map is also given in Table 4 for completeness. Numbers of articles appearing in each map, along with the number of ID/term pairs, citations, and/or similarity pairs (where known) used to generate the map, are also provided.

There are some issues around using terms to generate domain maps, two of which we address here. First is the distinction between the use of words (single words), compound terms (typically noun phrases parsed from titles or abstracts and consisting of multiple words), and specific terms (terms as they appear in lists from bibliographic sources). In this analysis, we have used specific terms, which can be either single words or multi-word phrases. For instance, in the ARIST data, the terms *SCIENTOMETRICS*, *SCIENTOMETRIC INDICATORS*, and *STATIONARY SCIENTOMETRIC INDICATORS* all occur multiple times. Multi-word terms tend to be more specific, and thus might be expected to produce clusters based on specifics. Parsing to single

words from the small number of specific compound terms is an option that leads to a large number of more general terms that can be used to index documents.

Table 4. Comparison of document maps.

	A: Carto-SOM	**B: ET-map**	**C: Co-term**	**D: LSA**	**Citation**
Basis	id/term	id/term	cosine	id/title words	direct/cocite
Years	1991-2001	1991-2001	1991-2001	1977-2001	1977-2001
# papers	1446	1286	1446	2702	1626
# ID/term pairs	5202	5202	5202		
# citations					4632
# sim pairs			72664	48435	18258

A second issue concerns the number of terms associated with each article. For the ARIST data, 34% of the articles with terms have 1-3 terms, 42% of the articles have 4-7 terms, 17% of the articles have 8-10 terms, and the remaining 7% have 11 or more terms. Chung and Lee (Chung & Lee, 2001) showed that cosine measures tend to emphasize high-frequency terms. Andre Skupin's experience (see section 6.1.3) with SOM is that articles with only 1 or 2 terms tend to either congregate in less dense areas of the SOM or in the middle of clusters, rather than at their edges.[29] Boyack's experience using a cosine co-term similarity with FDP is that articles with only 1 or 2 terms tend to be evenly distributed if the associated terms are specific, multi-word terms, but lie in the middle of large clusters if the associated terms are single word, general terms (e.g, *SCIENCE*).[30] Different results may arise from different layout routines; nevertheless, a definitive study on the effects of term types, the generality of terms, and term distributions is needed.

The term-based maps all show only a portion (around 50%) of the 2767 articles in the ARIST set. This is due to the lack of keywords for papers published prior to 1991, and to the lack of terms on 26% of the papers published since 1991. A total of 872 unique terms occurred more than once in the list of 5202 id/term pairs.

The citation map retained a few more papers (1626), but missed 1141 others because there were no citation links between them and any other member of the set. The LSA map, if the entire document-by-document were used (3.83 million similarity values above the matrix diagonal), would have retained all 2767 articles. However, due to the size of the matrix, only similarities of 0.9 or greater were used in ordination, and thus 65 articles were not retained in the map.

Figure 14 compares the layouts of the four term or word based maps. There is much that is similar between these four layouts. In all four, citation analysis and bibliometrics (yellow and green dots) are found together, while visualization (magenta) and semantics (blue) are mostly by themselves. There are some areas in each map with some overlap between semantics and visualization, and some very small regions with overlap between all four areas. Thus, at a macro

[29] Personal communication.
[30] Unpublished data.

level, the different layout techniques seem to give similar groupings from the term or word expressions. We have not done a neighborhood analysis to verify this statistically, but leave that to another time and study.

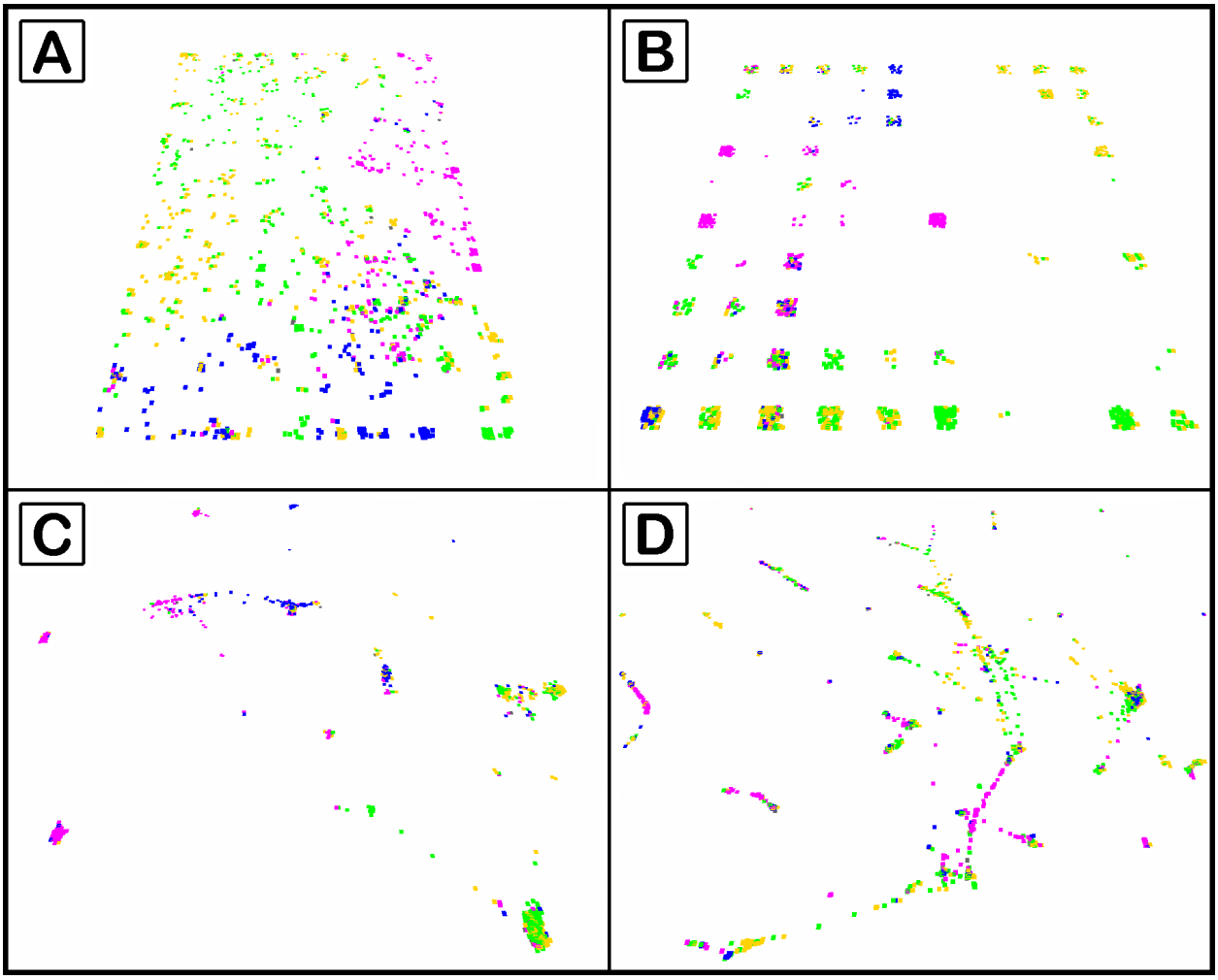


Figure 14. Comparison of layouts of four different document maps based on terms or words. A: Cartographic-SOM (compare Figure 10), B: ET-Map (compare Figure 9), C: Co-term (compare Figure 12, left), D: LSA (compare Figure 12, right). Dot color legend – YELLOW: citation analysis, GREEN: bibliometrics, BLUE: semantics, MAGENTA: visualization.

Figure 14 also makes clear the obvious visual differences between the layouts. The Cartographic SOM and ET-Maps each have more clusters than the co-term map (likely due to the parameters under which they were generated), while the LSA map shows more continuous structures. The three term-based layouts are shown in Figure 15 with strong co-term linkages (based on the cosine similarity) represented as lines. This view highlights more differences between the layouts. Intracluster linkages criss-cross the maps for the Cartographic-SOM and ET-Map, while there are few intercluster linkage trails for the co-term map. This suggests that, while the SOM-based methods break the data into more clusters, and tend to fill the space much more uniformly,

perhaps fewer clusters and less efficient space-filling can be justified. The LSA map is not compared here since title words and terms give different levels of information (Qin, 2000).

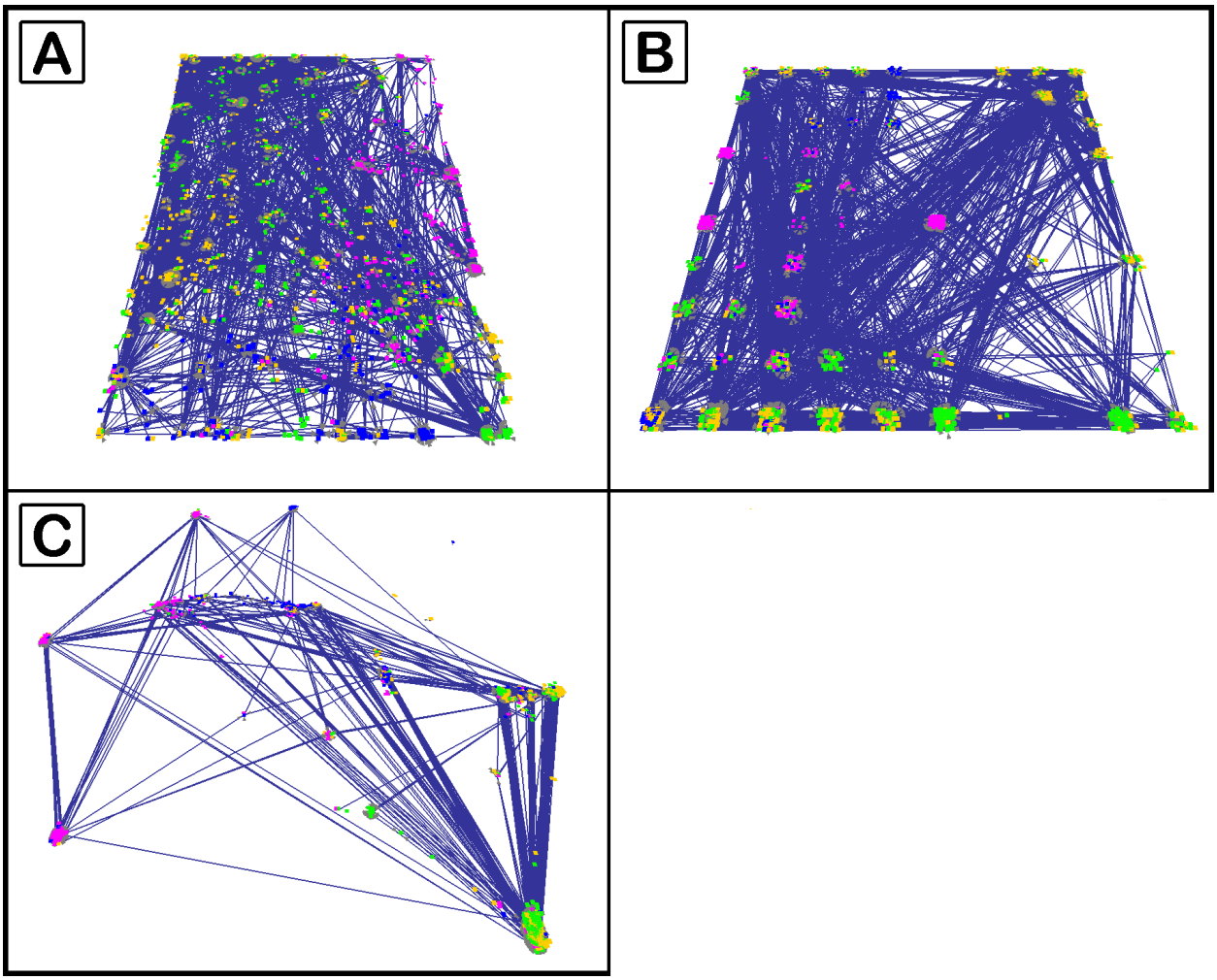


Figure 15. Strong co-term linkages based on cosine similarity for the three term-based document maps of Figure 14.

We now provide a comparison of reference-based citation maps with maps based on similarities between terms or words. The area around cluster D of Figure 11 seemed to be of interest due to the overlap between citation analysis, bibliometrics, and visualization. Browsing of that cluster showed that the visualization papers were all related to visualization of neural networks, and that they appeared in that cluster with a large number of citation and bibliometrics studies on decision-support systems. Browsing near the cluster also revealed that different types of citation analysis seemed to be clustered by analysis type.

The first panel of Figure 16 (citation map) shows the results of three queries to abstracts: author co-citation, co-word, and co-citation [NOT author co-citation]. The results of those queries lie in the clusters labeled A, B, and C, respectively, with very little scatter. Indeed the citation map portion of Figure 16 shows only a small portion of the overall citation map (compare Figure 11).

Browsing of individual articles confirmed that each cluster was comprised of articles where researchers used the technique central to the cluster. Identical queries were made to the Cartographic-SOM, co-term, and LSA maps, with the results shown in Figure 16. In these three maps, the query results are spread throughout the maps and are not associated with any discrete clusters. Thus, articles are clustered much better by technique when using references than when using term or title-based similarity measures. This is likely due to lack of consistent use of terms corresponding to techniques in either keyword lists or titles. We expect that maps based on similarities between words in abstracts would do a much better job of clustering by technique than do any of the measures shown here.

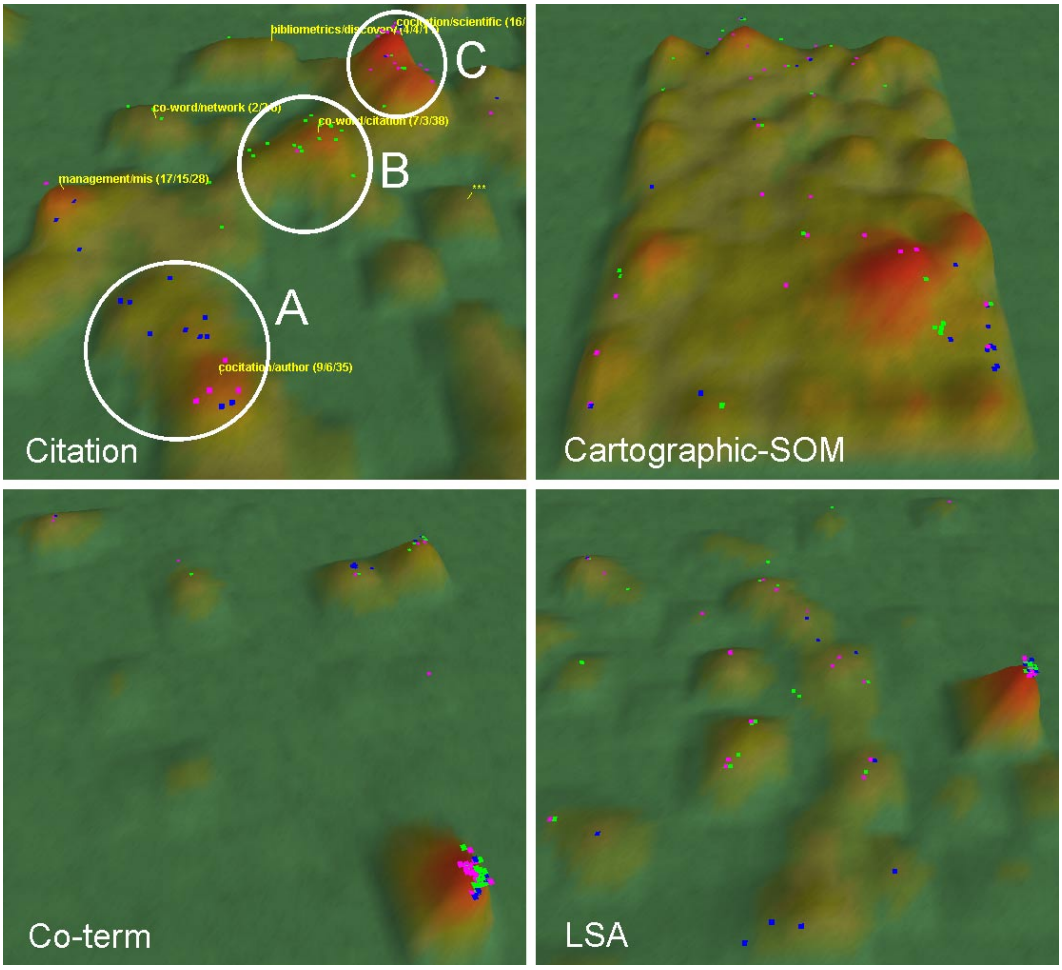


Figure 16. Comparison of distinct fields on the citation map with their counterparts on the term or title-based maps. Legend – BLUE: author co-citation (ACA), GREEN – co-word analysis and Leximappe, MAGENTA: co-citation analysis.

The analysis and layout comparisons presented here do not show that any one type of similarity method and layout are better than any other for producing domain visualizations. Rather, they show that trade-offs are involved and that the researcher should use the combination of similarity and layout techniques that are likely to aid in answering the questions at hand. Each researcher will, of course, have his or her favorite methods. We encourage all who are involved in domain visualization to broaden their horizons and expand the suite of methods that they use, to the

benefit of all who read and rely upon their work. Note also that this survey is largely derived from a quantitative approach. It is also important to take into account qualitative views.

7 PROMISING AVENUES OF RESEARCH

While working on this chapter the authors discussed a large and diverse number of potential ways to improve the generation of domain visualizations and their interpretation. In particular, we find the following to be of great interest:

- Ways to increase the accessibility of domain visualizations for non-experts.
- Utilizing domain visualization to help answer real-world questions.
- Bringing together leading researchers in different fields that contribute to the visualization of knowledge domains to improve the dissemination of results.
- Development of more robust, scalable algorithms.

Promising avenues of research that address those interests are discussed below.

Increasing the accessibility of domain visualizations for non-experts. Despite advances in visualization research, many non-expert users find the use of visualization tools to be unfamiliar and non-intuitive. Domain visualizations could greatly benefit from the incorporation of advanced visual perception (Ware, 2000); (Palmer, 1999) and cognitive principles into tools to aid the non-expert.

In addition, it seems to be desirable and advantageous to compare existing and novel algorithms on existing data sets and to contrast the results with human data, see (K. Börner, 2000) for first results. Ultimately, visualizations that best fit cognitive user models will be easier to understand and use.

Cartographic design considerations and specific techniques can enrich the design of domain visualizations in a number of ways. The value of using a geographic metaphor was first discussed in (Wise et al., 1995). Today, several geographers are involved in extending past work on geographic metaphors and primitives (Couclelis, 1998; Golledge, 1995) towards the development and testing of specific interfaces for spatialized browsing of non-geographic information (Fabrikant, 2000; Fabrikant & Buttenfield, 2001; Skupin & Buttenfield, 1996). Cartographic perspectives on information visualization are presented in (Skupin, 2000). The relevance of data structures and analytical tools common in geographic information systems is being investigated in (Skupin, 2001).

Utilizing domain visualization to help answer real-world questions. We believe that visualizations of knowledge domains can help to assess scientific frontiers, to forecast research vitality, to identify disruptive events/technologies/changes, and to find knowledge carriers, etc. For example, (Schwechheimer & Winterhager, 1999, 2001) applied a co-citation analysis based method to identify and analyze highly dynamic, rapidly developing research fronts of climate research. They used journal profiles, co-citation maps, and actor profiles as information elements. Results by Nederhof and Noyons (1992) indicate that bibliometric assessment of research performance is potentially useful in humanities disciplines. Much of the work done with VxInsight has been for competitive intelligence purposes (Boyack et al., 2002). Multi-SOM maps have been proposed for a knowledge-oriented analysis on science and technology via knowledge indicators (Polanco et al., 2001).

New commercial applications are also being aimed at domain analysis. VantagePoint[31] reads bibliographic data files from many different sources and automates such techniques as cross-correlation analyses and factor analyses for purposes of technology assessment and opportunities analysis. Other products, such as SemioMap[32] or Knowledgist[33], use linguistic techniques, semantic analysis, Bayesian models, or ontologies to understand the content of unstructured textual sources, whether in local files or on the Internet. These products seek to replace the "hunt-and-peck" method of keyword search by the ability to browse the entire search space for relevant documents. Internet Cartographer[34] is Inventix Software's solution to information overload on the Internet. It combines advanced artificial intelligence techniques with sophisticated information visualization techniques to build maps of accessed documents, organized in a hierarchy of over 500 pre-defined categories.

Recent work on so called "small world graphs" aiming at a graph theoretical analysis of the Web may turn out to be applicable to analyze and visualize bibliographic data and research networks as well. Work by Jon Kleinberg at Cornell University, and Prabhakar Raghavan and Sridhar Rajagopalan at IBM Almaden Research Center on "hubs" (documents which cite many other documents) and "authorities" (documents which are highly cited) (Kleinberg, 1999) could be used to identify excellent review articles and high quality papers respectively. It may also lead to improved measures to identify emerging research themes and communities based on authors of documents sharing some common theme.

Bringing together leading researchers. We believe that research on visualizing knowledge domains could be sped up considerably if one or two well-understood and expert-verified domain analysis data sets, e.g. in the form of a TREC data set[35], were made available for general use to rate different algorithms and visualization techniques. In addition, a centralized repository of data analysis and visualization software applicable to create domain visualizations would improve the dissemination of algorithmic knowledge, enable comparisons of algorithms, and save the time spent for re-implementing algorithms (Börner & Zhou, 2001). Assuming that ownership and privacy issues can be resolved, we believe that a data set and software repository would also boost creativity by easing access to existing work; consultation with others working on related topics; implementation of new (commercial) applications, which, in turn challenge the development and improvement of the algorithms; exploration of new ideas; and, last but not least, the dissemination of results to science (Shneiderman, 2000). The design of collaborative information visualizations that can be explored by multiple, potentially distributed users at the same time is also expected to improve collaborative data collection, access, examination, and management.

Development of more robust, scalable algorithms. More robust semantic similarity measures and highly scalable ordination techniques are needed. Most current similarity generation and layout algorithms require hours or more to produce results. This has not been a deterrent to

[31] http://www.thevantagepoint.com/ and http://tpac.gatech.edu/
[32] http://www.semio.com/
[33] http://www.invention-machine.com/
[34] http://www.inventix.com/
[35] http://trec.nist.gov/data.html

domain researchers who spend a great deal of time analyzing the results of domain maps, but will prove a deterrent to a future generation of users who want quick answers from small literature sets (500 or fewer articles) that they can download in real-time. Examples of this new generation of systems are AuthorLink and ConceptLink[36] by Xia Lin and colleagues, JAIR Space[37] by Mark Foltz, or Star Walker[38] by Chaomei Chen.

Layout algorithms that give robust answers are also needed. By robust answers, we mean a layout that does not change significantly with slight or even modest perturbations to the input. Recent work by Davidson (Davidson et al., 2001) using VxOrd showed that some clusters will break up with the introduction of small amounts of noise (random changes to similarities of 2% or less), while other clusters retain their structure with the addition of large amounts of noise (order 10%). In cases where robustness cannot be achieved, analysis to quantify the robustness of individual clusters can be done to aid researchers in knowing how much confidence to place in their results.

In the same way that the Web continually sprouts new pages, there is also a steady stream of new work in the scientific literature. Hence, research by Barabasi and his colleagues (Barabási & Albert, 1999; Barabási, Albert, & Jeong, 2000) on the development of algorithms that mimic the growth of the Web may be able to model the growth of scientific disciplines as well. Work on incremental ordination and layout algorithms is essential to visualize continually changing data. One advantageous feature of such algorithms is that they enable incremental update while preserving the main topology of the layout. Organic information design was first proposed by (Mackinlay, Rao, & Card, 1995) and applied recently by (Fry, 2000). It borrows ideas like growth, atrophy, responsiveness, homeostasis, and metabolism, and it applies them to design information visualization that one "can 'feed' data to, and watch how the data is digested." (Dodge, 2001).

Generative Topographic Mapping (GTM) (Bishop, Svensen, & Williams, 1998) is an alternative to SOMs in that it generates topographic maps. The model was developed at the Neural Computing Research Group, Aston University, UK[39] and is a novel form of latent variable modeling, which allows general non-linear transformations from latent space to data space. For the purposes of data visualization, the mapping is then inverted using Bayes' theorem, resulting in a posterior distribution in latent space. GTPs overcome most limitations of the self-organizing maps and might turn out to be a valuable method for the visualization of knowledge domains.

The genomics and bioinformatics world has been the home to many algorithmic innovations in recent years, many of which could be applied to data analysis, clustering, and visualization. Research focus is on matching algorithms, scalability of clustering methods, effectiveness of methods for clustering complex shapes and types of data, high-dimensional clustering techniques, and methods for clustering mixed numerical and categorical data in large databases. Continuing research into eigenvector and matrix techniques, and parallel algorithms may also lead to advances in information science.

[36] http://cite.cis.drexel.edu/
[37] http://www.infoarch.ai.mit.edu/jair/
[38] http://www.brunel.ac.uk/~cssrccc2/
[39] Papers, reports, software, demos etc. are available at http://www.ncrg.aston.ac.uk/GTM/

In the 1960s and early 1970s Belver Griffith applied bibliometrics and behavioral measures to reveal disciplinary communication structures. Recent work by Sandstrom (2001) shows that universal principles such as prey-choice models from optimal foraging theory – developed by biologists in the 1970s – can be successfully applied in the bibliographic microhabitat to explain information seeking and use behavior. Sandstrom (1994) was among the very first to see scholars as subsistence foragers in a socioecological framework. Research by Pirolli and Card (1999) supports the claim that foraging theory can be extended to understand information foraging and the evolution of knowledge domains, and to improve their visualizations.

Lastly we agree with Hjorland & Albrechtsen (1995) that information science in general and the visualization of knowledge domain in particular should be seen as a social rather than purely mental research area. This new view stresses the social, ecological, and purpose-oriented production and usage of knowledge.

8 CONCLUSIONS

Thank you for joining us on a long journey about the visualization of knowledge domains. We've gone through some history, the general process of generating domain maps, commonly used units and measures, specific techniques, and the application of several techniques to generate and compare diverse maps of the subject domain.

Guided through these maps you learned that this research field is currently divided into a few major islands, some of which – e.g., information visualization and semantics islands – are isolated. However, there are some interesting connecting points that should be exploited in future work.

We hope that our survey provides a starting point for researchers and practitioners to appreciate the richness and complexity of this fast evolving research field, to determine their own positions, to identify related work in other research areas, and to plan (interdisciplinary) collaborations and future work on promising applications.

We believe that research aimed at visualizing knowledge domains can benefit from importing and incorporating research in other fields as identified in section 7 to vastly improve the readability and effectiveness of domain visualizations. It can also contribute to the development of science in general by exporting methods and approaches to identify related work by experts in relevant research areas, assess research vitality, identify evolving research fields, etc. We hope that the various snapshots and our interpretations of this dynamic and interdisciplinary field of study produced in this grand tour can lead to insights into the essence of the field as a whole and its promising future.

9 ACKNOWLEDGEMENTS

The composition of a review article in today's flood of information can only be done in the middle of a tightly knit network of domain experts and colleagues supporting and connecting it to relevant research. We greatly appreciate the time and effort Bin Zhu, Hsinchun Chen, and André Skupin put into the generation, discussion and comparison of the ET-Map and

Cartographic SOM map. The section on SOM maps benefited from André Skupin's detailed feedback. We wish to thank Katherine W. McCain, Blaise Cronin, Henry Small, Pamela Sandstrom, and the anonymous reviewers for their very insightful comments. Ben Shneiderman and Alan Porter commented on an earlier version of this chapter.

We gratefully acknowledge support for this work by The Council for Museums Archives and Libraries in the UK (RE/089), Laboratory Directed Research and Development, Sandia National Laboratories, U.S. Department of Energy (DE-AC04-94AL85000), and an NIH/NIA demonstration fund for Mapping Aging Research.

10 BIBLIOGRAPHY

Ballay, J. M. (1994). *Designing Workscape: An interdisciplinary experience.* Paper presented at the CHI'94 Conference, Boston, MA: ACM Press, pp. 10-15.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, 509-512

Barabási, A.-L., Albert, R., & Jeong, H. (2000). Scale-free characteristics of random networks: The topology of the World Wide Web. *Physica, 281*, 69-77

Barbará, D., & Chen, P. (2000, August 20 - 23). *Using the fractal dimension to cluster datasets.* Paper presented at the Sixth ACM SIGKDD Conference on Knowledge Discovery in Data Mining, Boston, MA USA, pp. 260-264.

Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods: Theory and Applicatioms*. New York: John Wiley & Sons.

Bassecoulard, E., & Zitt, M. (1999). Indicators in a research institute: A multi-level classification of journals. *Scientometrics, 44*, 323-345

Battista, G., Eades, P., Tamassia, R., & Tollis, I. G. (1994). Algorithms for drawing graphs: An annotated bibliography. *Computational Geometry: Theory and Applications, 4*(5), 235-282

Battista, G. D., Eades, P., Tamassia, R., & Tollis, I. G. (1999). *Graph Drawing: Algorithms for the Visualization of Graphs*: Prentice Hall.

Bederson, B. B., Hollan, J. D., Perlin, K., Meyer, J., Bacon, D., & Furnas, G. (1996). Pad++: A zoomable graphical sketchpad for exploring alternate interface physics. *Journal of Visual Languages and Computing, 7*(1), 3-31. Available: <Go to ISI>://A1996UF41800002.

Bhattacharya, S., & Basu, P. K. (1998). Mapping a research area at the micro level using co-word analysis. *Scientometrics, 43*(3), 359-372

Bishop, C. M., Svensen, M., & Williams, C. K. I. (1998). GTM: The generative topographic mapping. *Neural Computation, 10*(1), 215-234. Available: <Go to ISI>://000071446700010.

Borg, I., & Groenen, P. (1996). *Modern Multidimensional Scaling*: Springer.

Borgman, C. L. (2000). Scholarly communication and bibliometrics revisited. In B. Cronin & H. B. Atkins (Eds.), *Web of Knowledge: A Festschrift in honor of Eugene Garfield* (pp. 143-162). Medford, NJ: Information Today.

Borgman, C. L. (Ed.). (1990). *Scholarly Communication and Bibliometrics*. Newbury Park, CA: Sage Publications.

Borgman, C. L., & Furner, J. (2002). Scholarly Communication and Bibliometrics. In B. Cronin & R. Shaw (Eds.), *Annual Review of Information Science and Technology*. Medford, NY: Information Today.

Börner, K. (2000, June 2-7). *Extracting and visualizing semantic structures in retrieval results for browsing.* Paper presented at the ACM Digital Libraries, San Antonio, Texas, pp. 234-235.

Börner, K. (2000, 19 -21July). *Searching for the perfect match: A comparison of free sorting results for images by human subjects and by Latent Semantic Analysis.* Paper presented at the Information Visualisation 2000, Symposium on Digital Libraries, London, England, pp. 192-197.

Börner, K., Dillon, A., & Dolinsky, M. (2000). *LVis - Digital Library Visualizer.* Paper presented at the Information Visualisation 2000, Symposium on Digital Libraries, London, England, pp. 77-81.

Börner, K., & Zhou, Y. (2001, July 25-27). *A Software Repository for Education and Research in Information Visualization.* Paper presented at the Information Visualisation Conference, London, England: IEEE Press, pp. 257-262.

Boyack, K. W., Wylie, B. N., & Davidson, G. S. (2002). Domain visualization using VxInsight for science and technology management. *Journal of the American Society for Information Science and Technology, 53*(9), 764-774

Boyack, K. W., Wylie, B. N., Davidson, G. S., & Johnson, D. K. (2000). *Analysis of patent databases using VxInsight.* Paper presented at the New Paradigms in Information Visualization and Manipulation '00, McLean, VA: ACM,

Bush, V. (1945). As we may think. *The Atlantic Monthly, 176*(1), 101-108

Cahlik, T. (2000). Comparison of the maps of science. *Scientometrics, 49*, 373-387

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From Translation to Network : The Co-Word Analysis. *Scientometrics, 5*(1), 78-78

Card, S., Mackinlay, J., & Shneiderman, B. (Eds.). (1999). *Readings in Information Visualization: Using Vision to Think*: Morgan Kaufmann.

Card, S. K. (1996). Visualizing retrieved information: A survey. *IEEE Computer Graphics and Applications, 16(2)*, 63-67

Chalmers, M. (1992, June 1992). *BEAD: Explorations in information visualisation.* Paper presented at the SIGIR '92, Copenhagen, Denmark: ACM Press, pp. 330-337.

Chen, C. (1997a). *Structuring and visualising the WWW with Generalised Similarity Analysis.* Paper presented at the the 8th ACM Conference on Hypertext (Hypertext '97), Southampton, UK: ACM Press, pp. 177-186.

Chen, C. (1997b). Tracking latent domain structures: An integration of Pathfinder and Latent Semantic Analysis. *AI & Society, 11*(1-2), 48-62

Chen, C. (1998a). Bridging the gap: The use of Pathfinder networks in visual navigation. *Journal of Visual Languages and Computing, 9*(3), 267-286

Chen, C. (1998b). Generalised Similarity Analysis and Pathfinder Network Scaling. *Interacting with Computers, 10*(2), 107-128

Chen, C. (1999a). *Information Visualisation and Virtual Environments*. London: Springer Verlag.

Chen, C. (1999b). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management, 35*(2), 401-420

Chen, C. (2002). *Mapping Scientific Frontiers*. London: Springer-Verlag.

Chen, C., & Carr, L. (1999, February 1999). *Trailblazing the literature of hypertext: Author co-citation analysis (1989-1998).* Paper presented at the the 10th ACM Conference on Hypertext (Hypertext '99), Darmstadt, Germany: ACM Press, pp. 51-60.

Chen, C., & Rada, R. (1996). Modelling situated actions in collaborative hypertext databases. *Journal of Computer-Mediated Communication, 2*(3)

Chen, C., Thomas, L., Cole, J., & Chennawasin, C. (1999). Representing the semantics of virtual spaces. *IEEE Multimedia, 6*(2)

Chen, C. M., & Paul, R. J. (2001). Visualizing a knowledge domain's intellectual structure. *Computer, 34*(3), 65-71. Available: <Go to ISI>://000167305100023.

Chen, C. M., Paul, R. J., & O'Keefe, B. (2001). Fitting the jigsaw of citation: Information visualization in domain analysis. *Journal of the American Society for Information Science and Technology, 52*(4), 315-330. Available: <Go to ISI>://000167208800005.

Chen, H. C., Houston, A. L., Sewell, R. R., & Schatz, B. R. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science, 49*(7), 582-603. Available: <Go to ISI>://000073044700002.

Chen, H. C., & Lynch, K. J. (1992). Automatic Construction of Networks of Concepts Characterizing Document Databases. *Ieee Transactions on Systems Man and Cybernetics, 22*(5), 885-902. Available: <Go to ISI>://A1992KE00400003.

Chen, H. C., Schuffels, C., & Orwig, R. (1996). Internet categorization and search: A self-organizing approach. *Journal of Visual Communication and Image Representation, 7*(1), 88-102

Chung, Y. M., & Lee, J. Y. (2001). A corpus-based approach to comparative evaluation of statistical term association measures. *Journal of the American Society for Information Science and Technology, 52*(4), 283-296. Available: <Go to ISI>://000167208800002.

Conklin, J. (1987). Hypertext: An introduction and survey. *IEEE Computer, September, 1987*, 17-41

Couclelis, H. (1998). Worlds of Information: The Geographic Metaphor in the Visualization of Complex Information. *Cartography and Geographic Information Science, 25*(4), 209-220

Crane, D. (1972). Invisible Colleges: Diffusion of Knowledge in Scientific Communities. Chicago: The University of Chicago Press.

Cronin, B., & Atkins, H. B. E. (2000). *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*: ASIST.
Davidson, G. S., Hendrickson, B., Johnson, D. K., Meyers, C. E., & Wylie, B. N. (1998). Knowledge mining with VxInsight: Discovery through interaction. *Journal of Intelligent Information Systems, 11*(3), 259-285
Davidson, G. S., Wylie, B. N., & Boyack, K. W. (2001). Cluster stability and the use of noise in interpretation of clustering. *Proc. IEEE Information Visualization 2001*, 23-30
Deboeck, G., & Kohonen, T. (Eds.). (1998). *Visual Explorations in Finance with Self-Organizing Maps*: London: Springer-Verlag.
Deerwester, S., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(6), 391-407
DeLooze, M. A., & Lemarie, J. (1997). Corpus relevance through co-word analysis: An application to plant proteins. *Scientometrics, 39*(3), 267-280
Ding, Y., Chowdhury, G., & Foo, S. (1999). Mapping the intellectual structure of information retrieval studies: an author co-citation analysis ; 1987-1997. *Journal of Information Science, 25*(1), 67-78
Ding, Y., Chowdhury, G. G., & Foo, S. (2000). Journal as markers of intellectual space: Journal co-citation analysis of information Retrieval area ; 1987-1997. *Scientometrics, 47*(1), 55-73
Dodge, M. (2001). Mapping how people use a website. *Mappa Mundi Magazine*. Available: http://mappa.mundi.net/maps/maps_022/#ref_4.
Dodge, M., & Kitchin, R. (2000). *Mapping Cyberspace*: Routledge.
Dourish, P., & Chalmers, M. (1994). *Running out of space: Models of information navigation.* Paper presented at the HCI '94,
Fabrikant, S. I. (2000). Spatialized Browsing in Large Data Archives. *Transactions in GIS, 4*(1), 65-78
Fabrikant, S. I., & Buttenfield, B. P. (2001). Formalizing Semantic Spaces for Information Access. *Annals of the Association of American Cartigraphers, 91*(2), 263-280
Feng, Y., & Börner, K. (2002, January 20-25). *Using Semantic Treemaps to Categorize and Visualize Bookmark Files.* Paper presented at the Visualization and Data Analysis. Proceedings of SPIE, P. C. C. Robert F. Erbacher, Matti Grohn, Jonathan C. Roberts, Craig M. Wittenbrink (Ed.) San Jose, CA, pp. 218-227.
Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph Drawing by Force-Directed Placement. *Software-Practice & Experience, 21*(11), 1129-1164
Fry, B. (2000). *Organic information design.* MIT, Master's Thesis.
Furnas, G. W. (1986). *Generalized fisheye views.* Paper presented at the CHI '86: ACM Press, pp. 16-23.
Furnas, G. W., & Zhang, X. (1998). *MuSE: a multiscale editor.* Paper presented at the Proceedings of the 11th annual ACM symposium on User interface software and technology, San Francisco, CA, USA, pp. 107 - 116.
Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science, 122*(108-111)
Garfield, E. (1994). Scientography: Mapping the tracks of science. *Current Contents: Social & Behavioural Sciences, 7*(45), 5-10
Garfield, E., Sher, I. H., & Torpie, R. J. (1964). *The use of citation data in writing the history of science*. Philadelphia: Institute for Scientific Information.
Garfield, E., & Small, H. (1989). *Identifying the changing frontiers of science*, [WWW]. The S. Neaman Press. Available: http://www.garfield.library.upenn.edu/papers/362/362.html [2000, June 26].
Gershon, N., Eick, S. G., & Card, S. (1998). Design: Information Visualization. *Interactions, 5*(2), 9-15
Gershon, N., & Ward, P. (2001). What Storytelling Can Do for Information Visualization. *Communications of the ACM, 44*(8). Available: http://www.acm.org/cacm/0801/0801toc.html.
Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics, 51*, 69-115
Glänzel, W., & DeLange, C. (1997). Modelling and measuring multilateral co-authorship in international scientific collaboration. Part II. A comparative study on the extent and change of international scientific collaboration links. *Scientometrics, 40*, 605-626
Golledge, R. G. (1995). Primitives of Spatial Knowledge. In T. L. Nyerges & D. M. Mark & R. Laurini & M. J. Egenhofer (Eds.), *Cognitive Aspects of Human-Computer Interaction for Geographic Information Systems* (pp. 29-44): Dordrecht: Kluwer.
Gorsuch, R. L. (1983). *Factor analysis* (2 ed.). Hillsdale, NJ: Erlbaum.
Halasz, F. (1988). Reflections on NoteCards: Seven issues for the next generation of hypermedia systems. *Communications of the ACM, 31*(7), 836-852

Halasz, F., Moran, T., & Trigg, R. (1986, December 1986). *Notecards in a nutshell.* Paper presented at the the ACM CHI+GI Conference, Austin,
Halliday, M. A. K., & Hasan, R. (1976). *Cohesion In English*: Longman.
Han, J., & Kamber, M. (2000). *Data Mining: Concepts and Techniques* (1st edition ed.): Morgan Kaufmann Publishers.
Harman, H. H. (1976). *Modern Factor Analysis*: University of Chicago Press.
Harter, S. P., Nisonger, T. E., & Weng, A. W. (1993). Semantic Relationships between Cited and Citing Articles in Library and Information-Science Journals. *Journal of the American Society for Information Science, 44*(9), 543-552. Available: <Go to ISI>://A1993LX51500004.
He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends, 48*, 133-159
Hearst, M. A. (1999). User interfaces and visualization. In R. Baeza-Yates & B. Ribeiro-Neto (Eds.), *Modern Information Retrieval* (pp. 257-224): Addison-Wesley.
Hendley, R. J., Drew, N. S., Wood, A. M., & Beale, R. (1995, October 30-31, 1995). *Narcissus: Visualizing information.* Paper presented at the Information Visualization '95 Symposium, Atlanta, GA: IEEE, pp. 90-96.
Herman, I., Melançon, G., & Marshall, M. S. (2000). Graph visualization and navigation in information visualization: a survey. *IEEE Transactions on Visualization and Computer Graphics*(1), 24-44. Available: http://www.cwi.nl/~ivan/AboutMe/CV/Recent References.html.
Hetzler, B., Whitney, P., Martucci, L., & Thomas, J. (1998). *Multi-faceted insight through interoperable visual information analysis paradigms.* Paper presented at the IEEE Information Visualization '98: IEEE,
Hjorland, B. (1997). *Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science*. Westport: Greenwood Press.
Hjorland, B., & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain analysis. *Journal of the American Society for Information Science, 46*(6), 400-425
Hollan, J. D., Bederson, B. B., & Helfman, J. (1997). Information visualization. In M. G. Helenader & T. K. Landauer & P. Prabhu (Eds.), *The Handbook of Human Computer Interaction* (pp. 33-48). The Netherlands: Elsevier Science.
ISI. (1981). *ISI atlas of science: Biochemistry and molecular biology, 1978/80*. Philadelphia, PA: Institute for Scientific Information.
Kaski, S., Honkela, T., Lagus, K., & Kohonen, T. (1998). WEBSOM - Self-organizing maps of document collections. *Neurocomputing, 21*(1-3), 101-117
Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*, 10-25
Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N., & Davidson, G. S. (2001). A Gene Expression Map for Caenorhabditis elegans. *Science, 293*, 2087-2092
King, J. (1987). A Review of Bibliometric and Other Science Indicators and Their Role in Research Evaluation. *Journal of Information Science, 13*(5), 261-276. Available: <Go to ISI>://A1987K402700001.
Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM, 46*(5), 604-632
Kohonen, T. (1985). *The self-organizing map.* Paper presented at the Proc. IEEE, vol. 73, pp. 1551-1558.
Kohonen, T. (1995). *Self-Organizing Maps*: Springer.
Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., & Saarela, A. (2000). Self organization of a massive document collection. *IEEE Transactions on Neural Networks, 11*(3), 574-585
Koike, H. (1993). The role of another spatial dimension in software visualization. *ACM Transactions on Information Systems, 11*(3), 266-286
Koike, H., & Yoshihara, H. (1993). *Fractal Approaches for Visualizing Huge Hierarchies.* Paper presented at the Proceedings of the 1993 IEEE Symposium on Visual Languages: IEEE/CS, pp. 55-60.
Kostoff, R. N., Eberhart, H. J., & Toothman, D. R. (1998). Database tomography for technical intelligence: A roadmap of the near-earth space science and technology literature. *Information Processing & Management, 34*(1), 69-85
Kruskal, J., B. (1964). Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika, 29*(2), 115-129.
Kruskal, J. B. (1964). Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika, 29*, 1-27
Kruskal, J. B. (1977). Multidimensional scaling and other methods for discovering structure. In K. Enslein & A. Ralston & H. Wilf (Eds.), *Statistical Methods for Digital Computers*. New York: Wiley.
Kruskal, J. B., & Wish, M. (1984). *Multidimensional Scaling*. Beverly Hills and London: Sage Publications.

Lamping, J., Rao, R., & Pirolli, P. (1995). *A focus+context technique based on hyperbolic geometry for visualizing large hierarchies.* Paper presented at the ACM CHI'95 Conference on Human Factors in Computing Systems, pp. 401-408.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284. Available: http://lsa.colorado.edu/.

Lawrence, S., Giles, C. L., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer, 32*(6), 67-71. Available: http://www.neci.nj.nec.com/home.

Lee, R. C. T., Slagle, J. R., & Blum, H. (1977). A triangulation method for the sequential mapping of points from N-Space to Two-Space. *IEEE Transactions on Computer*(26), 288-292

Leydesdorff, L. (1994). The generation of aggregated journal-journal citation maps on the basis of the CD-ROM version of the Science Citation Index. *Scientometrics, 47*, 143-164

Leydesdorff, L., & Wouters, P. (2000). *Between texts and contexts: Advances in theories of citation*. Available: http://www.chem.uva.nl/sts/loet/citation/rejoin.htm [2000, June 26].

Lin, X. (1997). Map displays for information retrieval. *Journal of the American Society for Information Science, 48*(1), 40-54

Lin, X., Soergel, D., & Marchionini, G. (1991). *A self-organizing semantic map for information retrieval.* Paper presented at the 14th. Annual International ACM/SIGIR Conference on Research & Design in Information Retrieval, Y. C. A. Brokstein, G. Salton, & V.V.Raghuvan (Ed.): New York: ACM Press, pp. 262-269.

Lin, Y., & Kaid, L. L. (2000). Fragmentation of the intellectual structure of political communication study: Some empirical evidence. *Scientometrics, 47*(1), 143-164

Mackinlay, J. D., Rao, R., & Card, S. K. (1995). *An Organic User Interface For Searching Citation Links.* Paper presented at the CHI, Denver, Colorado: ACM,

Mahlck, P., & Persson, O. (2000). Socio-bibliometric mapping of intra-departmental networks. *Scientometrics, 49*(1), 81-91

Mandelbrot, B. B. (1988). *Fractal Geometry of Nature*: W H Freeman & Co.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science, 41*(6), 433-443

McCain, K. W. (1995). The structure of biotechnology research-and-development. *Scientometrics, 32*(2), 153-175

McCain, K. W. (1998). Neural networks research in context: A longitudinal journal cogitation analysis of an emerging interdisciplinary field. *Scientometrics, 41*(3), 389-410

McCormick, B. H., DeFanti, T. A., & Brown, M. D. (1987). *Visualization in scientific computing*: Report of the NSF Advisory Panel on Graphics, Image Processing and Workstations.

McQuaid, M. J., Ong, T. H., Chen, H. C., & Nunamaker, J. F. (1999). Multidimensional scaling for group memory visualization. *Decision Support Systems, 27*(1-2), 163-176. Available: <Go to ISI>://000084396700010.

Mostafa, J., Quiroga, L. M., & Palakal, M. (1998). Filtering medical documents using automated and human classification methods. *Journal of the American Society for Information Science, 49*(14), 1304-1318. Available: <Go to ISI>://000077119700008.

Mukherjea, S. (1999). Information visualization for hypermedia systems. *ACM Computing Surveys*

Munzner, T. (1997). *H3: Laying out large directed graphs in 3D hyperbolic space.* Paper presented at the the 1997 IEEE Symposium on Information Visualization, J. Dill & N. Gershon (Eds.), Phoenix, AZ: IEEE, pp. 2-10.

Munzner, T. (1998). Exploring large graphs in 3D hyperbolic space. *IEEE Computer Graphics and Applications, 18*(4), 18-23. Available: http://graphics.stanford.edu/papers/h3cga/html/.

Narin, F., & Moll, J. K. (1977). Bibliometrics. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 12, pp. 35-58). White Plains, NY: Knowledge Industry Publications.

Nederhof, A. J., & Noyons, E. C. M. (1992). International comparison of departments research performance in the humanities. *Journal of the American Society for Information Science, 43*(3), 249-256. Available: <Go to ISI>://A1992HK01700006.

Nederhof, A. J., & Zwaan, R. A. (1991). Quality Judgments of Journals as Indicators of Research Performance in the Humanities and the Social and Behavioral-Sciences. *Journal of the American Society for Information Science, 42*(5), 332-340. Available: <Go to ISI>://A1991FN78000003.

Newman, M. E. J. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E, 64*, 016131

Newman, M. E. J. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E, 64*, 016132

Nisonger, T. E., Harter, S. P., & Weng, A. (1992). Subject Relationships between Cited and Citing Documents in Library and Information-Science. *Proceedings of the Asis Annual Meeting, 29*, 13-19. Available: <Go to ISI>://A1992JV32500004.

Noyons, E. (2001). Bibliometric mapping of science in a science policy context. *Scientometrics, 50*(1), 83-98. Available: <Go to ISI>://000167170600007.

Noyons, E. C. M., Moed, H. F., & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the American Society for Information Science, 50*(2), 115-131. Available: <Go to ISI>://000078411700003.

Noyons, E. C. M., Moed, H. F., & Van Raan, A. F. J. (1999). Integrating research performance analysis and science mapping. *Scientometrics, 46*(3), 591-604. Available: <Go to ISI>://000084660600017.

Noyons, E. C. M., & Van Raan, A. F. J. (1998). Advanced mapping of science and technology. *Scientometrics, 41*(1-2), 61-67. Available: <Go to ISI>://000071770100006.

Noyons, E. C. M., & van Raan, A. F. J. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science, 49*(1), 68-81. Available: <Go to ISI>://000071048000009.

Palmer, S. E. (1999). *Vision science: From Photons to Phenomenology*. Cambridge, MA: Bradford Books/MIT Press.

Persson, O. (2001). All author citations versus first author citations. *Scientometrics, 50*(2), 339-344. Available: <Go to ISI>://000167664900010.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review, 106*(4), 643-675. Available: <Go to ISI>://000083338800001.

Pirolli, P., Schank, P., Hearst, M., & Diehl, C. (1996). *Scatter/Gather browsing communicates the topic structure of a very large text collection.* Paper presented at the the Conference on Human Factors in Computing Systems (CHI '96), Vancouver, BC: ACM Press,

Polanco, X., Francois, C., & Lamirel, J. C. (2001). Using artificial neural networks for mapping of science and technology: A multi-self-organizing-maps approach. *Scientometrics, 51*(1), 267-292

Price, D. D. (1961). *Science since Babylon*. New Haven: Yale University Press.

Price, D. D. (1963). *Little Science, Big Science*.Unpublished manuscript.

Price, D. D. (1965). Networks of scientific papers. *Science, 149*, 510-515

Price, D. D. (1986). *Little Science, Big Science and Beyond*. New York: Columbia University Press.

Qin, J. (2000). Semantic similarities between a keyword database and a controlled vocabulary database: An investigation in the antibiotic resistance literature. *Journal of the American Society for Information Science, 51*(2), 166-180

Raghavan, V. V., & Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science, 37*(5), 279-287

Raghupathi, W., & Nerur, S. P. (1999). Research themes and trends in artificial intelligence: An author co-citation analysis. *Intelligence, 10*(2)

Rennison, E. (1994). *Galaxy of News -- An Approach to Visualizing and Understanding Expansive News Landscapes.* Paper presented at the ACM Symposium on User Interface Software and Technology,

Rorvig, M. (1999). Images of similarity: a visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science, 50*(8), 639-651

Roweis, S. T., & Saul, L. K. (2000). Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science, 290*(5500), 2323-2326

Salton, G., Allan, J., & Buckley, C. (1994). Automatic Structuring and Retrieval of Large Text Files. *Communications of the Acm, 37*(2), 97-108. Available: <Go to ISI>://A1994MU59200013.

Salton, G., Allan, J., Buckley, C., & Singhal, A. (1994). Automatic-Analysis, Theme Generation, and Summary of Machine-Readable Texts. *Science, 264*(5164), 1421-1426. Available: <Go to ISI>://A1994NP22100032.

Salton, G., & Buckley, C. (1988). Parallel Text Search Methods. *Communications of the Acm, 31*(2), 202-215. Available: <Go to ISI>://A1988M209000010.

Salton, G., & Buckley, C. (1991). Global Text Matching for Information-Retrieval. *Science, 253*(5023), 1012-1015. Available: <Go to ISI>://A1991GC98200040.

Salton, G., Yang, C., & Wong, A. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613-620

Salvador, M. R., & Lopez-Martinez, R. E. (2000). Cognitive structure of research: Scientometric mapping in sintered materials. *Research Evaluation, 9*, 189-200

Sandstrom, P. E. (2001). Scholarly communication as a socioecological system. *Scientometrics, 51*(3), 573-605. Available: <Go to ISI>://000170653400011.
Sarkar, M., & Brown, M. H. (1994). Graphical fisheye views. *Communications of the ACM, 37*(12), 73-84
Schvaneveldt, R. W. (1990). *Pathfinder Associative Networks: Studies in Knowledge Organization. Norwood*. NJ: Ablex Publishing.
Schwechheimer, H., & Winterhager, M. (1999). Highly dynamic specialities in climate research. *Scientometrics, 44*(3), 547-560
Schwechheimer, H., & Winterhager, M. (2001). Mapping interdisciplinary research fronts in neuroscience: A bibliometric view to retrograde amnesia. *Scientometrics, 51*(1), 311-318
Shneiderman, B. (1992). Tree visualization with tree-maps: A 2-d space filling approach. *ACM Transactions on Graphics, 11*(1), 92 - 99
Shneiderman, B. (1996). *The eyes have it: a task by data type taxonomy for information visualizations.* Paper presented at the Symposium on Visual Languages, Boulder, CO: Proceedings of IEEE, pp. 336-343.
Shneiderman, B. (1997). *Designing the User Interface : Strategies for Effective Human-Computer Interaction* (3 ed.): Addison-Wesley Pub Co.
Shneiderman, B. (2000). Creating creativity: User interfaces for supporting innovation. *ACM Transactions on Computer-Human Interaction, 7*(1), 114-138
Skupin, A. (2000, October 2000). *From Metaphor to Method: Cartographic Perspectives on Information Visualization.* Paper presented at the Proceedings of InfoVis 2000, Salt Lake City, UT: IEEE Computer Society, pp. 91-97.
Skupin, A. (2001). *Cartographic Considerations for Map-Like Interfaces to Digital Libraries.* Paper presented at the Visual Interfaces to Digital Libraries - Its Past, Present, and Future, JCDL Workshop, K. Borner & C. Chen (Eds.), Roanoke, VA,
Skupin, A., & Buttenfield, B. P. (1996). *Spatial Metaphors for Visualizing Very Large Data Archives.* Paper presented at the GIS/LIS'96, Bethesda: American Society for Photogrammetry and Remote Sensing, pp. 607-617.
Small, H. (1994). A SCI-MAP case study: Building a map of AIDS research. *Scientometrics, 30*(1), 229-241
Small, H. (1995). Navigating the citation network. *Proc. 58th Annual Meeting of the American Society for Information Science, 32*, 118-126
Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics, 38*(2), 275-293
Small, H. (1999a). A passage through science: Crossing disciplinary boundaries. *Library Trends, 48*(1), 72-108
Small, H. (1999b). Visualizing science by citation mapping. *Journal of the American Society for Information Science, 50*(9), 799-813. Available: <Go to ISI>://000081198600009.
Small, H. (2000). Charting pathways through science: Exploring Garfield's vision of a unified index to science. *Web of Knowledge - a Festschrift in Honor of Eugene Garfield*, 449-473. Available: <Go to ISI>://000167603300024.
Spence, B. (2000). *Information Visualization*: Addison-Wesley.
Spence, B. (2001). *Information Visualization*: Addison-Wesley.
Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science, 290*(5500), 2319-2323
Thurstone, L. L. (1931). Multiple Factor Analysis. *Psychological Review, 38*, 406-427
Tryon, R. C. (1939). *Cluster Analysis*: New York: McGraw-Hill.
Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
Tufte, E. R. (1990). *Envisioning Information*. Cheshire, CT: Graphics Press.
Tufte, E. R. (1997). *Visual Explanations*. Cheshire, CT: Graphics Press.
van Dalen, H. P., & Henkens, K. (2001). What Makes a Scientific Article Influential? The Case of Demographers. *Scientometrics, 50*(3), 455-482
van Raan, A. (2000). The Pandora's box of citation analysis: Measuring scientific excellence - The last evil? In H. Atkins (Ed.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 301-319): ASIS.
Ware, C. (2000). *Information Visualization: Perception for Design*: Morgan Kaufmann.
White, H. D., Buzydlowsky, J., & Xia, L. (2000). *Co-cited author maps as interfaces to digital libraries.* Paper presented at the IEEE Information Visualization Conference, London, UK, pp. 25-30.
White, H. D., & Griffith, B. C. (1982). Authors and markers of intellectual space: Co-citation studies of science, technology and society. *Journal of Documentation, 38*(4), 255-272
White, H. D., & McCain, K. W. (1989). Bibliometrics. In M. E. Williams (Ed.), *Annual Review of Information Science & Technology* (Vol. 24, pp. 119-186). Amsterdam, The Netherlands: Elsevier Science.

White, H. D., & McCain, K. W. (1997). Visualization of literatures. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 32, pp. 99-168). Medford, NJ: Information Today.

White, H. D., & McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science, 49*(4), 327-356

Wilson, C. S. (2001). Informetrics. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology* (Vol. 34, pp. 107-286). Medford, NJ: Information Today.

Wise, J. A. (1999). The ecological approach to text visualization. *Journal of the American Society for Information Science, 50*(13), 1224-1233

Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995, October 30-31, 1995). *Visualizing the non-visual: Spatial analysis and interaction with information from text documents.* Paper presented at the EEE Symposium on Information Visualization '95, Atlanta, Georgia, USA: IEEE Computer Society Press,

Yang, C. C., Chen, H., & Hong, K. K. (1999, August 11 - 14). *Visualization tools for self-organizing maps.* Paper presented at the Fourth International ACM Conference on Digital Libraries, Berkeley, CA USA, pp. 258-259.

Zhu, D., & Porter, A. L. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. *Technological Forecasting and Social Change, to appear*