

Information Visualization, Human-Computer Interaction, and Cognitive Psychology: Domain Visualizations

Kevin W. Boyack[†], Brian N. Wylie, George S. Davidson

Sandia National Laboratories*
Albuquerque, NM 87185 USA
[†]505-844-7556

[†] kboyack@sandia.gov

ABSTRACT

Digital libraries stand to benefit from technology insertions from the fields of information visualization, human-computer interaction, and cognitive psychology, among others. However, the current state of interaction between these fields is not well understood. We use our knowledge visualization tool, VxInsight[®], to provide several domain visualizations of the overlap between these fields. Relevant articles were extracted from the Science Citation Indexes (SCI and Social SCI) using keyword searches. An article map, a semantic (co-term) map, and a co-author network have been generated from the data. Analysis reveals that while there are overlaps between fields, they are not substantial. However, the most recent work suggests areas where future collaboration could have a great impact on digital libraries of the future.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *clustering, relevance feedback*

H.5.2 [Information Interfaces and Presentation]: User Interfaces – *graphical user interface, interaction styles, screen design*

I.5.3 [Pattern Recognition]: Clustering – *similarity measures*

General Terms

Algorithms, Human Factors.

Keywords

Cognitive science, human-computer interaction, information visualization, digital libraries, interactive navigation, data mining, clustering, VxInsight.

1. INTRODUCTION

The amount of information becoming available in digital form is

increasing exponentially. Many institutions, while interested in providing digital information to their users, are only slowly making the shift from paper to digital libraries. There is a pronounced need for advances in techniques and tools to aid the individual user in finding and gleaning knowledge from relevant information.

It is felt by many researchers that such advances would be enhanced by insertion from the fields of information visualization, human-computer interaction, and cognitive psychology. Yet, to date, it is unclear how much interaction there has been between these fields, and how much impact each has had on advances in digital libraries.

The purpose of this paper is to explore the history and current state of overlap between these four fields (including digital libraries) by analysis of bibliographic information. We employ our visualization tool, VxInsight[®] [1], which was developed to build and explore maps of technology using data from the Science Citation Index (SCI). Over the past few years, we have found that VxInsight has broad application to mapping and navigation of many different types of data [2,3]. In this paper, we provide an overview of related work and tools, some background on the VxInsight tool and process, and several different visualizations of the domain comprised of the four fields mentioned. We close with a summary and suggestions for future work.

2. RELATED WORK

2.1 Literature maps

Various efforts to map the structure of science from literature have been undertaken for many years. The majority of these studies are performed at the discipline or specialty level. Maps are often based on similarity between journal articles using citation analysis [4], co-occurrence or co-classification using keywords, topics, or classification schemes [5,6], or journal citation patterns [7,8]. Latent semantic analysis (LSA) has been used to map papers based on co-occurrence of words (or authors) in titles, abstracts, or full text sources [9,10]. In addition, domain maps based on author co-citation analysis [11] are becoming more common. Many of these studies probe the dynamic nature of science, and the implications of the changes.

Presented at *User Interfaces to Digital Libraries – Its Past, Present, and Future*, a Workshop at *Joint Conference on Digital Libraries '01*, June 24-28, 2001, Roanoke, VA.

* 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.

Once a similarity matrix is defined, algorithms are used to cluster the data objects (e.g., articles or patents). Common clustering or ordination methods for producing maps include multidimensional scaling, hierarchical clustering, k-means algorithms, Pathfinder network scaling, and self-organizing maps. The standard mapping output for early literature studies was a circle plot where each cluster was represented by an appropriately sized circle. Links between circles provide relationship information. Traditionally, map outputs have been paper-based and only resolve structure at a few discrete levels. However, in recent years, several systems have been reported that use a computer display and allow some navigation, browsing, and filtering of the map space.

2.2 Visualization tools

SENTINEL [12] is a Harris Corporation package that combines a retrieval engine using n-grams and context vectors for effective query with the VisualEyes visualization system. The visualization tool allows the user to interact with document clusters in a three-dimensional space. Chen [9,13] uses a VRML 2.0 viewer in conjunction with Generalized Similarity Analysis to display authors (as spheres) and the Pathfinder linkage network based on author co-citation analysis. Citation rates are shown as the 3rd dimension in these VRML maps. Börner [10] uses the CAVE environment at Indiana University to interface with documents in a virtual library. Documents are clustered using latent semantic analysis. Features such as shape, color, and labeling are used to identify features of each document. Document details are available on demand through a hypertext link.

Self-organizing maps have been used in many venues, including the organization of document spaces [14]. This technique is used to position documents, and then display them in a two-dimensional contour-map-like display in which color represents density.

Two packages that are more similar to VxInsight are SCI-Map developed by ISI [15], and the SPIRE suite of tools which originated at Pacific Northwest National Laboratory [16,17]. SCI-Map uses a hierarchically-nested set of maps to display the document space at varying levels of detail. This nesting of maps allows drilling down to subsequent levels. Each map is similar to the traditional circle plot, where the size of the circle can indicate the density of documents contained in the circle, or some measure of importance. Relationships at each discrete level are indicated by links between circles.

Like VxInsight, SPIRE maps objects to a two-dimensional plane so that related objects are near each other, and provides tools to interact with the data. SPIRE has two visualization approaches. In the Galaxies view, documents are displayed as a scatter plot. This interface allows drilling down to smaller sections of the scatter plot, and provides some summarization tools. In the Themescape view, a high-level terrain display, similar to that in VxInsight, is used. Themescape visualizes specific themes as mountains and valleys, where the height of a mountain represents the strength of the theme in the document set.

3. VxINSIGHT DESCRIPTION

VxInsight® is a powerful and flexible PC-based tool for exploring data collections. It works by providing access to data in an intuitive visual format that is easy to interpret and that aids natural navigation. VxInsight exploits the human capability to visually detect patterns, trends, and relationships by presenting the data as

a landscape, a familiar representation that we are adept at interpreting, and which allows very large data sets to be represented in a memorable way.

Figure 1 shows the general process through which data must pass to produce a VxInsight® map. A typical database, represented as a spreadsheet in the figure, would consist of a few thousand objects (the rows), with one or more attributes arranged as tables (the columns). These must be processed to compute similarities for each pair of data elements, which are then used to construct an abstract graph. In this graph of nodes and arcs, the nodes represent individual data elements and the arcs are the similarities between the elements.

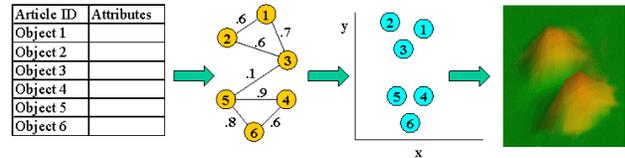


Figure 1. Data processed into a VxInsight map.

VxInsight uses a force-directed placement algorithm, VxOrd [3], to cluster data elements using pairwise similarity values as input. This algorithm uses a random walk technique, where the criteria for moving nodes is the minimization of energy given by:

$$E_{x,y} = \left[\sum_{i=0}^n (w_i \times l_i^2) \right] + D_{x,y}$$

where $E_{x,y}$ is the energy of a node with n edges at a specific x,y location, w_i are the pairwise similarities, l_i is the Euclidean distance between this node and the node connected by edge i , and $D_{x,y}$ is a density measure with respect to the area around point x,y . Use of a density field for the repulsive term is an order $O(N)$ process, thus ordination using VxOrd can be done rapidly. Another advantage of VxOrd is that the number, size, and position of clusters are dictated by the data, not by an arbitrary assignment of the number of clusters. The output from this clustering process is an x,y location on the abstract visualization surface for each data element.

In VxInsight, these coordinates are used to generate the mountain terrains. The height of each mountain is proportional to the number of objects beneath it. Labels for peaks are generated dynamically from any attribute in the database by showing the two most common words, phrases, numbers, etc. in a cluster for that attribute. This reveals the content of the objects that comprise each peak, and provides context for further navigation and query.

VxInsight supports multi-resolution zooming into the landscape to explore interesting regions in greater detail, which reveals structure on multiple scales. Following each mouse click, the landscape and labels are recalculated, to give a new, higher resolution view of the desired terrain. Temporal data can be viewed using a time slider to reveal growth and reduction in areas of interest, new emerging areas, and bridged regions that have merged together.

Data access and retrieval is achieved via an ODBC connection to the user's data source. Clicking on a single data element provides detail on demand for that item (such as author, source, title, etc.) A query window allows the user to interrogate the data source, resulting in colored markers on the terrain showing those items matching the query. The distribution of query markers in the

context of the terrain with its labels can be very meaningful to the analyst. Various analysis tasks can be accomplished by combining navigation, multiple queries, and time sliding functions.

4. DOMAIN VISUALIZATIONS

Several different visualizations were prepared to show the domain comprised by the fields of information visualization (IV), human-computer interaction (HCI), cognitive psychology (CP), and digital libraries (DL). One of the main advantages of domain visualization is the ability to combine and explore related work from different fields.

The first step in this process was to procure an appropriate set of bibliographic data. Data were retrieved from the Science Citation Indexes (SCI and Social SCI) through the SciSearch web interface used at Sandia. A short list of search terms related to the four fields was compiled (see Table 1) and was queried against titles, abstracts, and keywords for years 1991-present for the SCI, and 1995-present for the SSCI. The number of articles retrieved by each query is shown in Table 1, along with unique numbers of articles once duplicates are removed.

Table 1. Search terms and numbers of articles retrieved.

Search Term and Field		'91-'01		Unique	
		SCI	SSCI	SCI	SSCI
cognitive model	(CP)	363	430	363	270
cognitive science	(CP)	386	450	375	280
cognitive psychology	(CP)	236	415	216	289
cognitive system	(CP)	162	125	148	49
visualization	(IV)	434	51	434	14
exploration	(IV)	149	34	143	21
navigation	(IV)	53	14	50	6
browsing	(IV)	26	4	26	3
digital library	(DL)	246	219	238	127
human computer interface	(HCI)	565	234	539	72
human computer interaction	(HCI)	204	49	170	13
mental model		322	532	292	340
TOTALS		3146	2557	2994	1484

A total of 4478 unique articles were thus retrieved, with approximately 700, 800, 2000, 370 articles in the IV, HCI, CP, and DL fields, respectively. The majority of duplicate articles were the result of overlap between the SCI and SSCI, rather than overlap between search terms. The term 'mental models' was included in the original search, but was found to have little overlap with fields other than CP. Thus, it will not be discussed further.

Three different domain maps based on these data were produced: an article map, a semantic (keyword) map, and an author map. Each will be described further below.

4.1 Article Map

A map of articles in the IV/HCI/CP/DL domain was generated based on the number of ISI keywords (DE field in the SciSearch output) in common between each pair of articles. The similarity metric used (w_{ij}) is a cosine similarity given by the expression:

$$w_{ij} = T_{ij} / \sqrt{n_i * n_j}$$

where T_{ij} is the number of keywords in common for articles i and j , and n_k is the number of keywords for any article k . A threshold value for w_{ij} of 0.2 was applied to keep low w_{ij} values from

dominating the clustering. The highest w_{ij} value for each article whose maximum w_{ij} was below the threshold was also kept, in order to not exclude those articles from the map.

Clustering was performed using VxOrd, resulting in a map of 3142 articles. 1336 articles were not given positions on the map since they had no keywords in common with any other article. 60% of the DL articles are in this category, suggesting that additional text-based analyses (e.g. LSA) are needed to fully understand the DL overlaps. Figure 2 shows the domain map for three separate 2-year time periods.

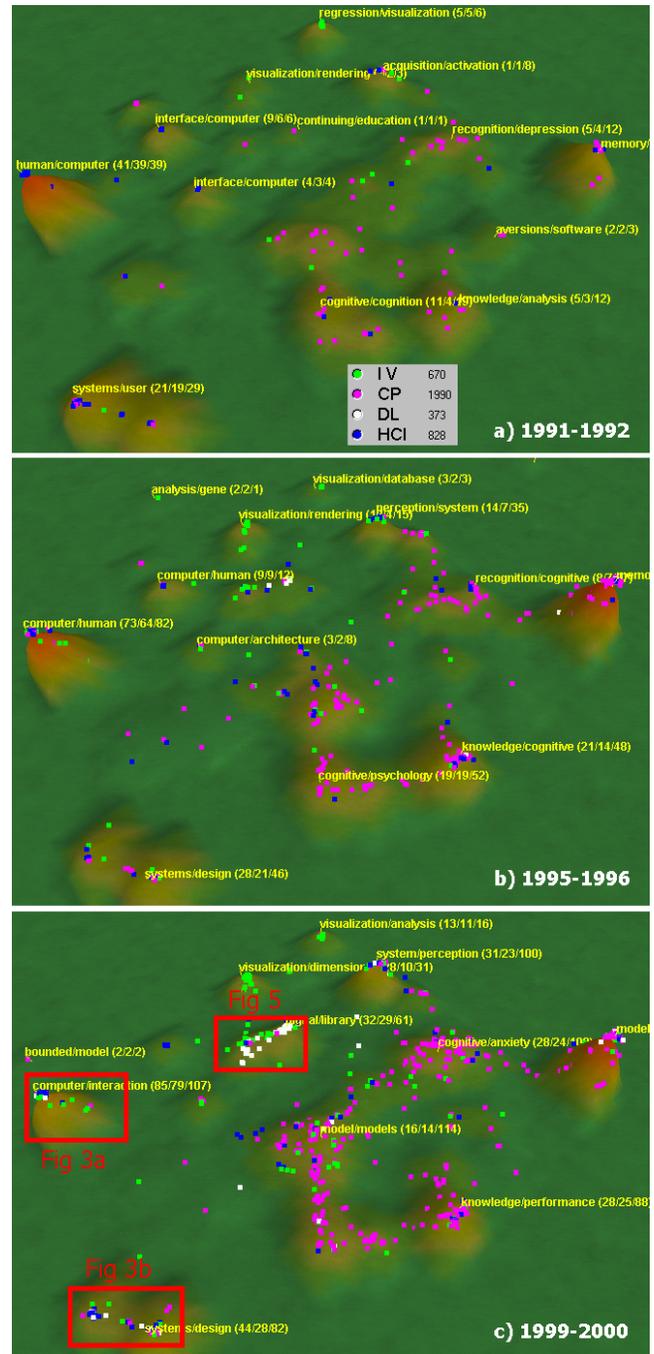


Figure 2. IV/HCI/CP/CL domain for three time periods.

An overview of Figure 2 reveals that although there are some overlaps in the four fields, they are not extensive. Peaks in the lower right portion of the terrain are dominated by CP (magenta dots), with few IV and HCI papers in some areas. IV work (green dots) is found near the top center of the terrain, and does show significant overlap with DL (white dots) in one peak. Most of the HCI work (blue dots) is in the peaks at the far left and lower left. These peaks show perhaps more overlap between the four fields than any others. Detailed views of these two HCI peaks are shown in Figure 3, which covers the entire time period from 1991-2000. The far left peak (3a) contains mostly HCI material with a scattering of IV and CP at the edges. The peak at the lower left (3b) is actually a ridge of two clusters with bridging material between them, where the left-most peak has more HCI material and the right-most peak has more CP material. Close examination of this structure indicates that the left-most cluster is concerned with interfaces and design, while the right-most cluster is more concerned with systems and cognition. The work in the center deals with relevance of HCI design.

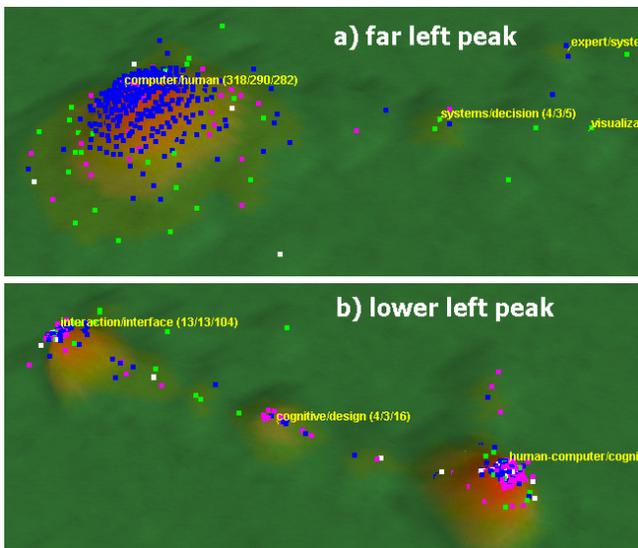


Figure 3. Detail on HCI peaks from Figure 2.

Figure 2 also shows trends in publishing in the four fields. DL work appears in the 1995-96 time frame, and grows through the 1999-2000 time frame. In addition, some DL work has moved from the core DL work in the center of the terrain to the HCI peaks by year 2000. This indicates that DL may be receiving some benefit or insertion from the HCI/CP work that forms the HCI clusters. This observation is tempered somewhat by the lack of many direct query overlaps shown in Figure 4. In fact, only 8 papers retrieved with the DL query (green dots in Figure 4) were also retrieved using any of the other queries. Thus, while DL work may be benefiting indirectly from the HCI/CP work, very little work connects them directly.

Additional trends from Figure 2 include a slight shift in HCI work (lower left peaks) from interface design to a system design with more cognitive modeling input. This is indicated by the shift in peak sizes and query marker distributions in Figure 2.

Perhaps the most exciting overlap from a DL standpoint occurs in the DL peak near the top center of Figure 2, which shows a significant overlap with IV. A detailed view of this region is shown in Figure 5. Here, *retrieval design* and *database retrieval*

are sandwiched between core DL and IV clusters. Several HCI and CP papers are also found in this region of convergence between IV and DL. This topic of *retrieval* is thus at the overlap of all four fields, and is an area in which DL can benefit from collaborations across the other three disciplines.

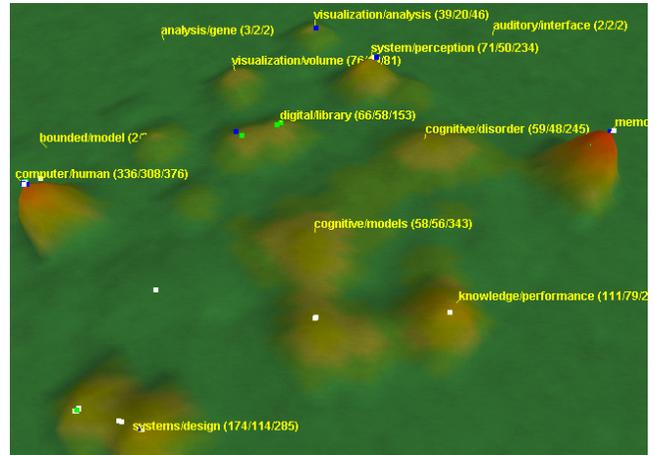


Figure 4. Papers overlapping fields based on search terms.

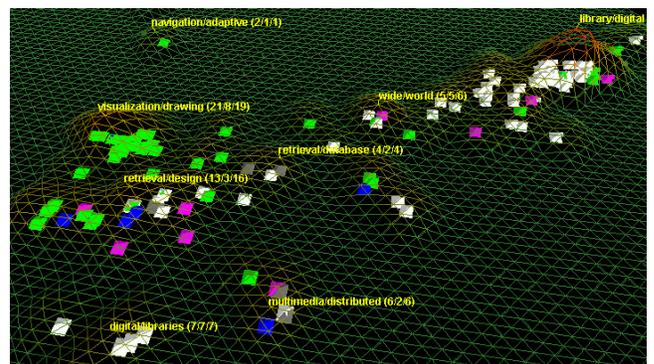


Figure 5. Detail on the DL / IV peak from Figure 2.

This conclusion that future work based on convergence between the IV, HCI, CP, and DL fields should focus on information retrieval may appear elementary. However, this analysis provides a formal basis to that conclusion, and may also suggest specific work that can be built upon for such studies. A list of the articles that appear in Figure 5 is too large for this paper, but may be obtained from the author.

4.2 Semantic Map

In addition to the article map described above, a semantic map (co-term) representing the domain was also created. This map was created using ISI keywords as terms. No words parsed from titles or abstracts were used as terms. A cosine similarity between terms i and j was used

$$w_{ij} = T_{ij} / \sqrt{n_i * n_j}$$

where T_{ij} is the number of co-occurrences of terms in articles i and j , and n_k is the number of occurrences for term k . This analysis was restricted to terms occurring at least twice in the corpus of documents, comprising 2373 terms in all. No similarity threshold was employed.

Figure 6 shows those terms in the semantic map that occur 30 times or more in the document corpus. The center map shows the

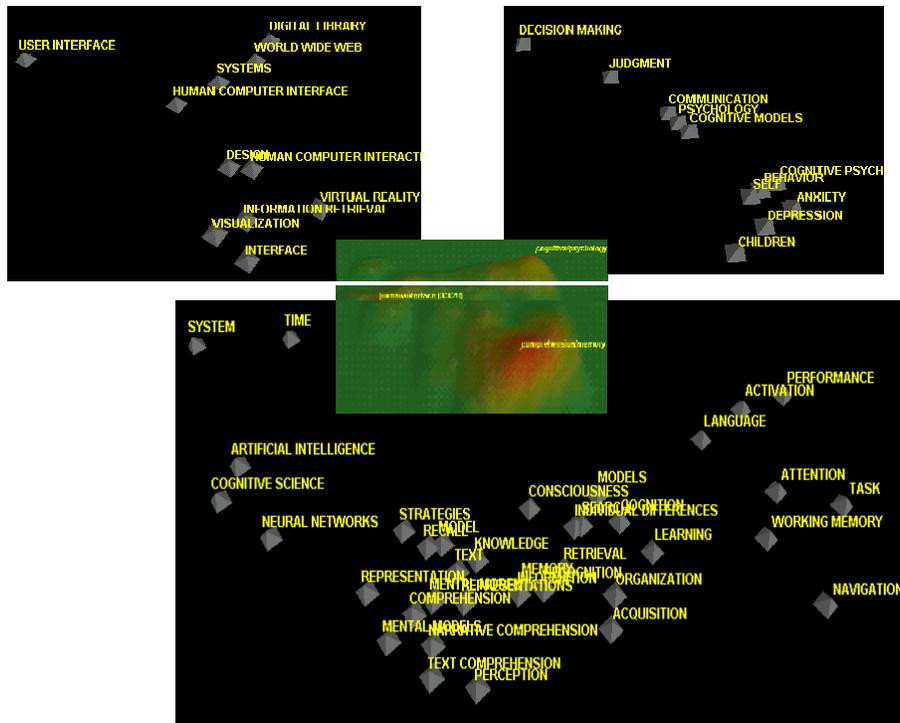


Figure 6. Semantic map for the IV/HCI/CP/DL domain.

spatial relationship between the clusters of terms, while the other three segments show individual terms and their spatial relationships.

The cluster at the upper left contains terms related to three of our four disciplines: IV, HCI, and DL. The term *information retrieval* is also there, which indicates its prominence, and which adds credibility to the conclusion reached above that it is a current topic of overlap between the four fields, and should be the focus of future work.

The cluster at the upper right is concerned with the medical side of cognitive psychology, while the large cluster at the bottom of the terrain focuses on various cognitive processes and information structures. Clearly, the bridge between the cognitive sciences and processes and the IV/HCI/DL fields is not as strong semantically as it needs to be to have a strong impact on the digital libraries of the future.

Figure 7 shows the changes over time in terms lying between the upper left and lower peaks from Figure 6. Terms such as *automatic analysis* and *graph algorithms* are bridging the space between the IV/HCI/DL terms and the cognitive sciences. Tools with these attributes may prove fruitful in inserting more of the benefits from cognitive sciences in digital libraries of the future.

4.3 Co-author Network

A co-author network for the IV/HCI/CP/DL domain has also been generated. Only authors with 2 or more papers in the document corpus were included. 885 authors matched this criterion. Association was once again calculated using a cosine similarity where the union term T_{ij} denotes the number of papers co-authored by a pair of authors.

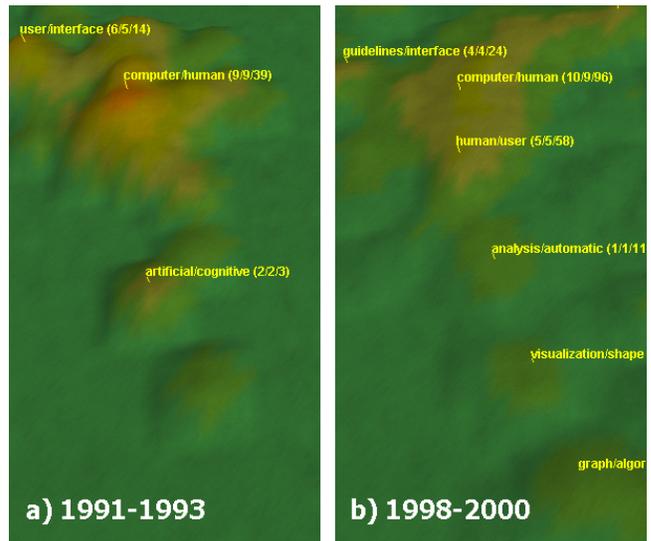


Figure 7. Changes in terms over time.

The co-author network is represented in Figure 8. Authors whose papers were retrieved by queries to the four fields are shown as dots of different colors. There are very few instances where dots of more than one color occur in the same local cluster. In addition, there are many, many clusters in this map, but only one arrow visible (at the scale indicated by the figure) that join more than one cluster. This indicates that there really is no co-author network established in this domain. Rather, the majority of researchers have interactions with a small group of others, doing research in one of the four fields. For a convergence in the IV,

HCI, CP, and DL fields to truly occur, much more collaboration across fields is needed.

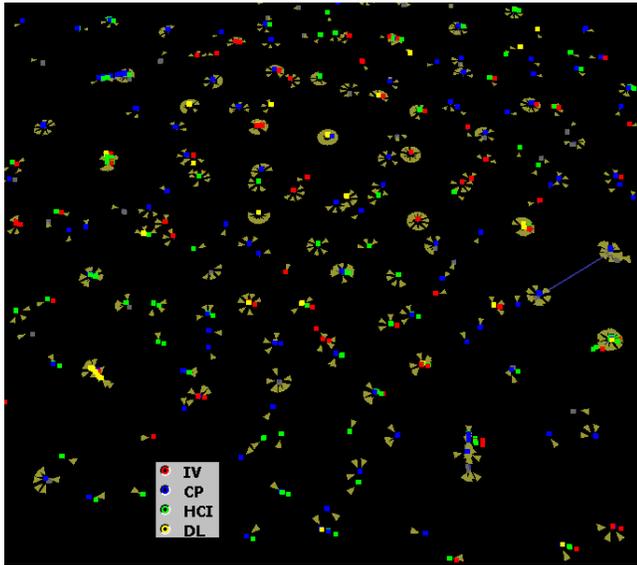


Figure 8. Co-author network. Arrows show connections.

5. SUMMARY AND FUTURE WORK

We have produced three visualizations of the domain comprised of the fields of information visualization, human-computer interaction, cognitive psychology, and digital libraries. Analysis based on dynamic views of the maps indicates that the current overlap between these fields is not substantial. However, the analyses also indicate that there are areas where recent research has occurred, and where future research should be focused to greatly benefit the digital libraries of the future.

Collaborative work between researchers in the four fields on the topics of *information retrieval*, and *advanced graphing and automated analysis algorithms* seem to be most fruitful at the moment. Such collaborations should not only benefit digital libraries, but other areas of focus between pairs of the disciplines.

Given that 60% of the DL articles had no keywords in common with any other article, and thus were left out of the article map of section 4.1, we advocate a more substantial textual analysis of the data, perhaps using LSA to extract words from titles and abstracts. If there are additional overlaps between DL and the other fields, they should be uncovered using those methods.

We also advocate a further analysis of authors across the four fields using an author co-citation analysis rather than the co-author study performed here. A co-citation analysis could point out a more extensive research network than we are led to believe exists based on the co-author network alone, and could suggest researchers and institutions that should collaborate more fully.

6. REFERENCES

[1] 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, 259-285.

[2] Boyack, K.W., Wylie, B.N., Davidson, G.S. & Johnson, D.K. (2000). Analysis of patent databases using VxInsight. *ACM*

New Paradigms in Information Visualization and Manipulation '00, McLean, VA, Nov. 10, 2000.

[3] Davidson, G.S., Wylie, B.N. & Boyack, K.W. (2001). Cluster stability and the use of noise in interpretation of clustering. Accepted by *IEEE Information Visualization '01*.

[4] Small, H. (1997). Update on science mapping: creating large document spaces. *Scientometrics* 38, 275-293.

[5] Noyons, E.C.M. & Van Raan, A.F.J. (1998). Advanced mapping of science and technology. *Scientometrics* 41, 61-67.

[6] Spasser, M.A. (1997). Mapping the terrain of pharmacy: co-classification analysis of the International Pharmaceutical Abstracts database. *Scientometrics* 39, 77-97.

[7] 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* 31, 59-84.

[8] Bassecoulard, E. & Zitt, M. (1999). Indicators in a research institute: A multi-level classification of scientific journals. *Scientometrics*, 44, 323-245.

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

[10] Börner, K. (2000). Extracting and visualizing semantic structures in retrieval results for browsing. *ACM Digital Libraries '00*, San Antonio, TX, June 2000.

[11] 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), 327-355.

[12] Fox, K.L., Frieder, O., Knepper, M.M. & Snowberg, E.J. (1999). SENTINEL: A multiple engine information retrieval and visualization system. *Journal of the American Society for Information Science* 50(7), 616-625.

[13] Chen, C., 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.

[14] Honkela, T., Kaski, S., Kohonen, T. & Lagus, K. (1998). Self-organizing maps of very large document collections: Justification for the WEBSOM method. In I. Balderjahn, R. Mathar & M. Schader (Eds.) *Classification, Data Analysis, and Data Highways*. Berlin: Springer.

[15] Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science* 50(9), 799-813.

[16] Hetzler, B., Whitney, P., Martucci, L., & Thomas, J. (1998). Multi-faceted insight through interoperable visual information analysis paradigms. *Proceedings of IEEE Information Visualization '98*, 137-144.

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