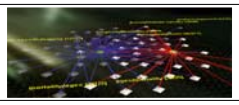
Mapping Interdisciplinary Research Domains

Katy Börner
School of Library and Information Science
INDIANA UNIVERSITY
BLOOMINGTON
katy@indiana.edu

Presentation at the Parmenides Center for the Study of Thinking, Island of Elba, Italy
June 26th, 2004

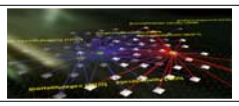


Overview

- Mapping Scientific Domains – Why?
- Maps of Scientific Domains – Samples
- Mapping Scientific Domains – How?

- Related Research Projects

- Infrastructure



Mapping Scientific Domains – Why?

To answer questions such as:

- What are the major research areas, experts, institutions, regions, nations, grants, publications, journals in xx research?
- Which areas are most insular?
- What are the main connections for each area?
- What is the relative speed of areas?
- Which areas are the most dynamic/static?
- What new research areas are evolving?
- Impact of xx research on other fields?
- How does funding influence the number and quality of publications?

Answers are needed by funding agencies, companies, and researchers.

Maps of Scientific Domains

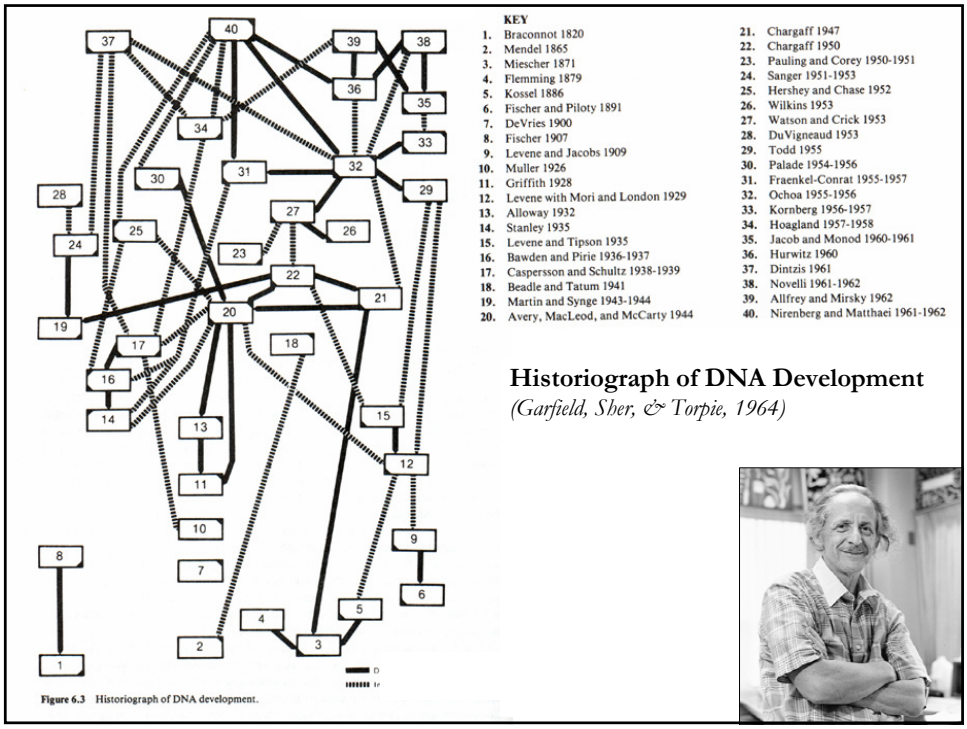
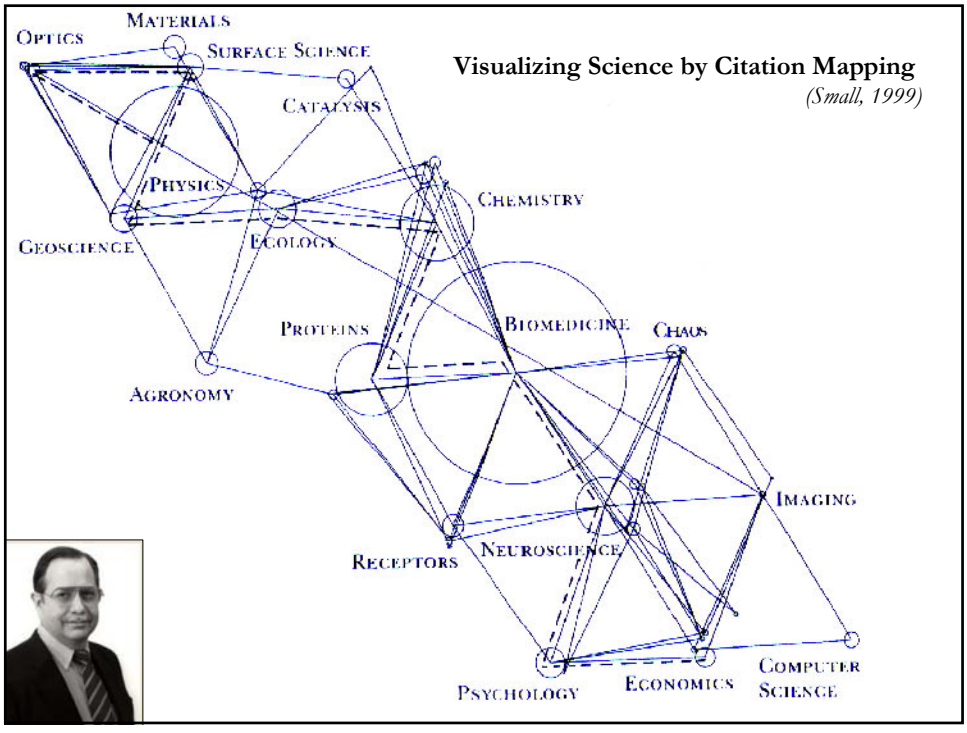


Figure 6.3 Historiograph of DNA development.

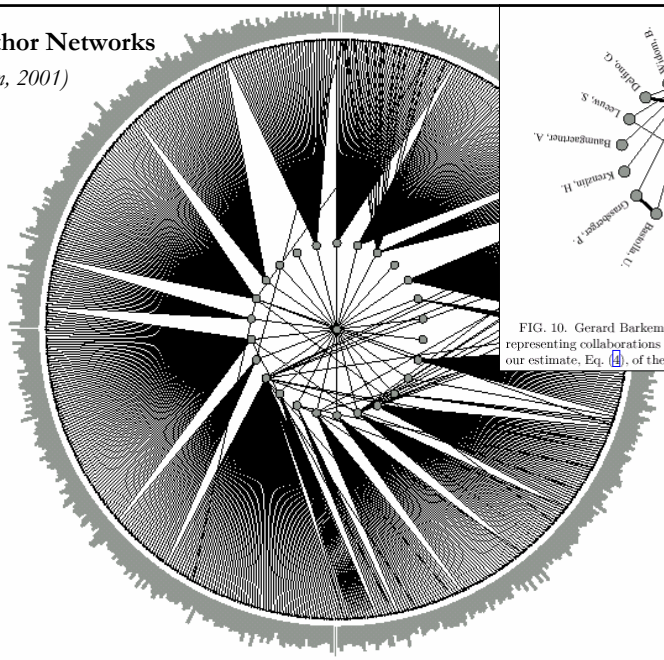
Historiograph of DNA Development
(Garfield, Sher, & Torpie, 1964)



Visualizing Science by Citation Mapping
(Small, 1999)

Co-author Networks

(Newman, 2001)



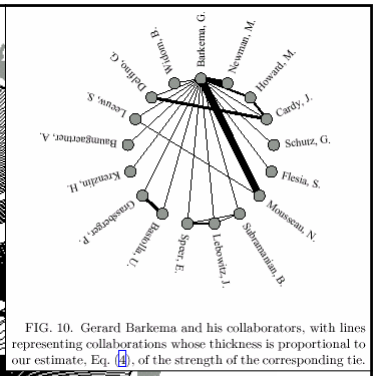
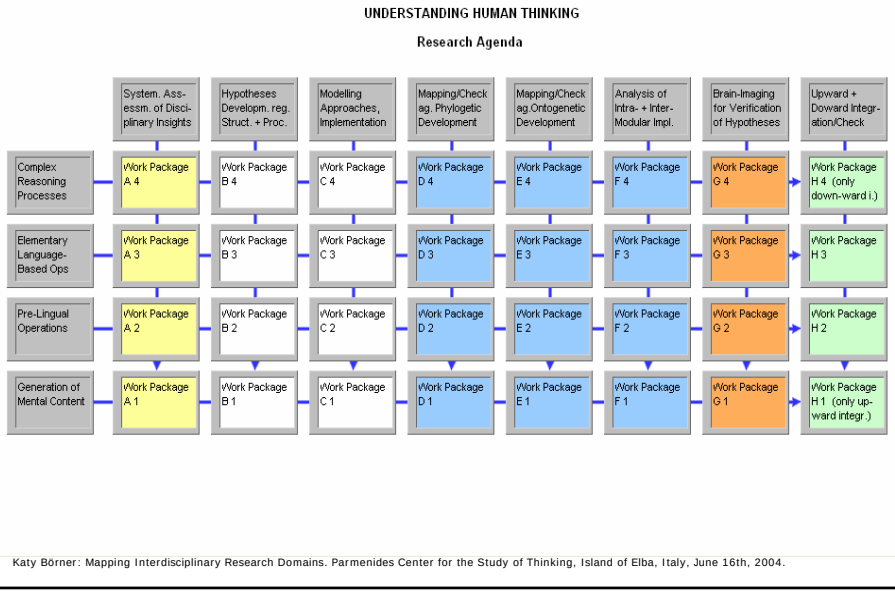
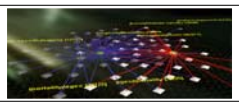


FIG. 10. Gerard Barkema and his collaborators, with lines representing collaborations whose thickness is proportional to our estimate, Eq. (4), of the strength of the corresponding tie.

FIG. 6. The point in the center of the figure represents the author of the paper you are reading, the first ring his collaborators, and the second ring their collaborators. Collaborative ties between members of the same ring, of which there are many, have been omitted from the figure for clarity.



Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, Italy, June 16th, 2004.



Mapping Scientific Domains – Why Visuals?

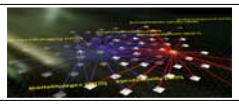
"Information Visualization is a process of transforming data and information that are not inherently spatial, into a visual form allowing the user to observe and understand the information."

(Source: Gershon and Eick, First Symposium on Information Visualization)

- Rooted in geograp
- Not even 15 years
- Far reaching (IR,
- Tremendous pote

Humans can detect a single dark pixel in a 500 x 500 array of white pixels in less than a second. This screen can be replaced every second by another, enabling a search of 15 million pixels in a minute (Ware, 2000).

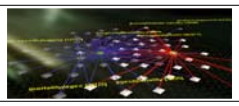
Also, people have a truly remarkable ability to recall pictorial images. In one study, Standing, Conezio, & Haber (1970) showed S's 2560 pictures, each for 10 seconds over 7 hours, in a 4-day period. Afterwards, S's were asked to classify pictures presented at a rate of 16 pictures/second and they achieved better than 90% accuracy.



Mapping Scientific Domains – How?

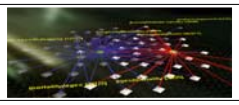
DATA EXTRACTION	UNIT OF ANALYSIS	MEASURES	LAYOUT (often one code does both similarity and ordination steps)		DISPLAY
			SIMILARITY	ORDINATION	
SEARCHES ISI INSPEC Eng Index Medline ResearchIndex Patents etc. BROADENING By citation By terms	COMMON CHOICES Journal Document Author Term	COUNTS/FREQUENCIES 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) LSA 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

Börner, Katy, Chen, Chaomei, and Boyack, Kevin. (2003) Visualizing Knowledge Domains. In Blaise Cronin (Ed.), *Annual Review of Information Science & Technology*, Volume 37, Medford, NJ: Information Today, Inc./American Society for Information Science and Technology, chapter 5, pp. 179-255.



Review Paper: Visualizing Knowledge Domains

Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, Italy, June 16th, 2004.



Review Paper: Visualizing Knowledge Domains

Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, Italy, June 16th, 2004.

SEARCH TERM USED	Number of matching articles
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

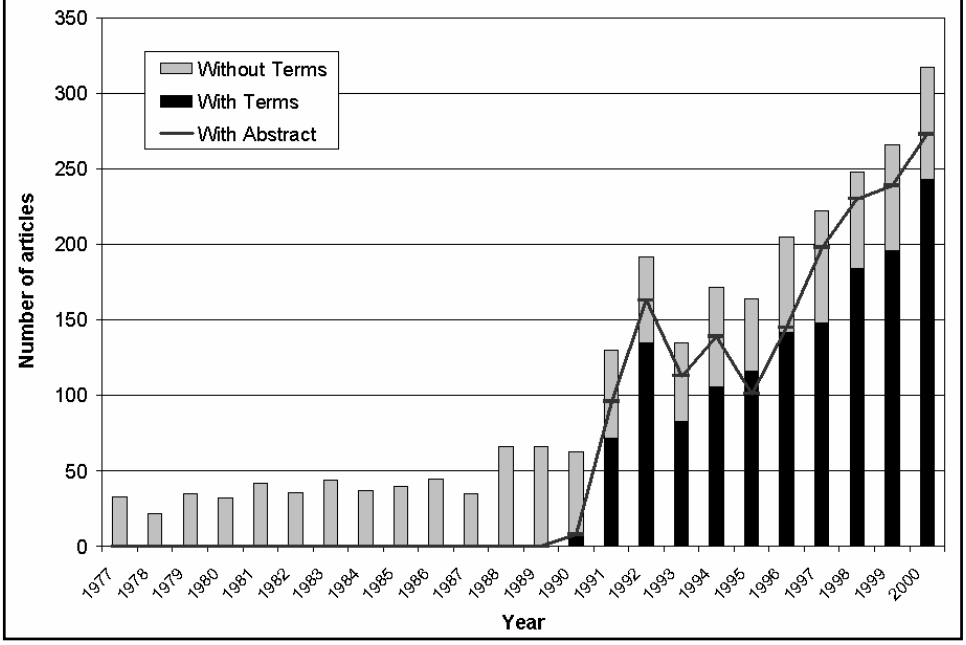
ARIST Data Set

Retrieved from Science Citation Index (SCI) and Social Science Citation Index (SSCI).

The 2764 unique articles match citation analysis, semantics, bibliometrics, visualization related terms in titles, abstracts, and terms for the years 1977-July 27, 2001.

Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, I

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



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

Citation counts dropped throughout the 90s.
Most recent articles are cited infrequently due to their young age.

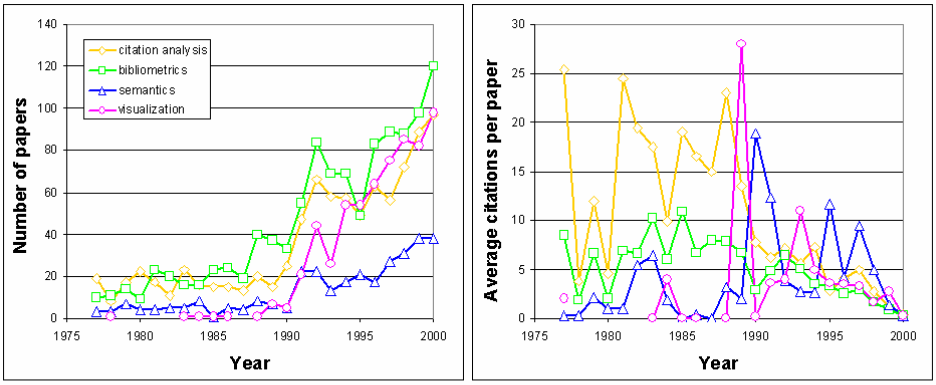


Figure 3. Numbers of articles by field per year in the ARIST data set with average citation counts. Articles contribute to counts in more than one field if retrieved by queries from multiple fields.

Figure shows dramatic increase in publishing in citation analysis and bibliometrics starting in the late 80s and the birth of IV around the same time.

The Importance of Good Data

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.

Limitations of the ARIST Data Set

- No abstracts or terms prior to 1991.
- Terms are available for only 71%. Abstracts are available for 81% of the articles published since 1991.
- Limited book, journal, conference coverage. No patents, policy changes, media coverage, Nobel prices, quality of graduate programs, …

Three different kinds of visualizations:

1. GSA/StarWalker use *Principal Component Analysis* to break down domain into components.
2. ET-Maps and Cartographic *Self Organizing Maps* display overall domain structure as adjacent regions.
3. VxInsight uses a modified *Force Directed Placement* algorithm named VxOrd to display a 'data landscape'.

The different visualizations provide different views of the domain and enable a comparison of

algorithms.

(1) GSA/StarWalker

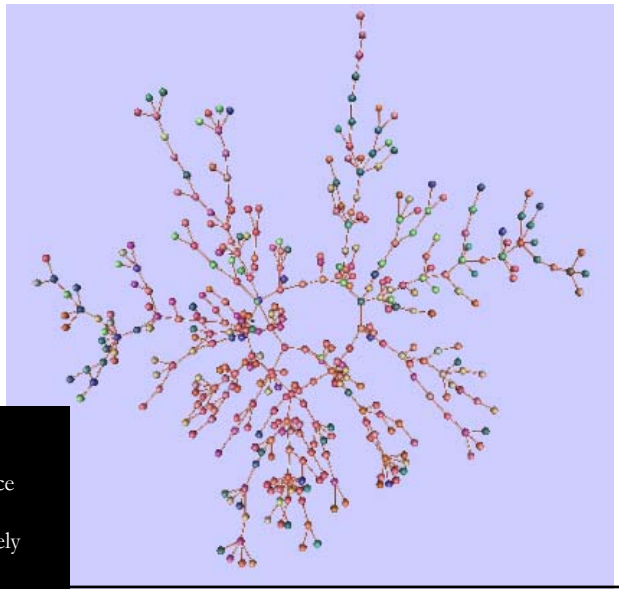
- Author co-citation analysis
- Document co-citation analysis

Procedure:

- Select a set of highly cited authors/documents (at least 10 citations).
- Compute co-citation frequencies.
- Apply Pathfinder Network Scaling to determine interconnectivity structure.
- Apply factor analysis to define intellectual groupings (e.g. mapping science, social studies of science, bibliometrics)
- Visualize and display citation impact factor atop the intellectual groupings.

Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, Italy, June 16th, 2004.

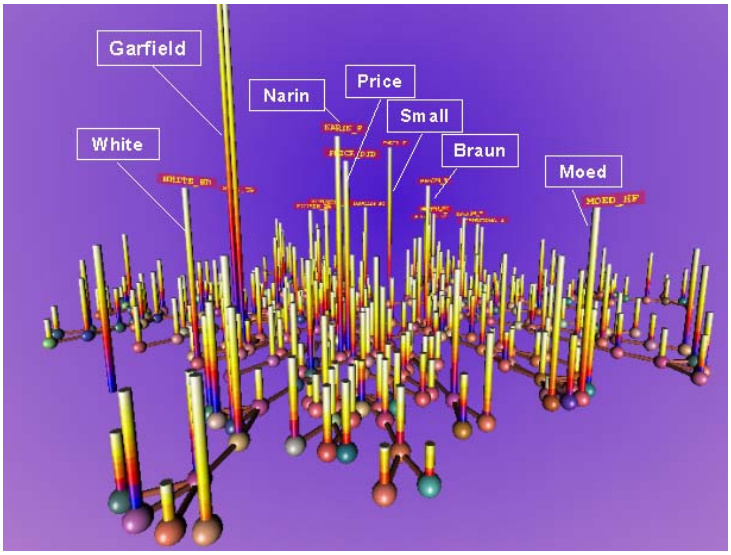
The Author Co-citation Map (1977-2001)
consists of 380 authors. The map is dominated by the largest specialty of citation indexing. No strong concentration of other specialties are found, which implies the diversity of the domain.



Color code:
red - mapping science
green – social studies of science
Blue – bibliometrics
The three factors cummulatively explain 63% of the variance

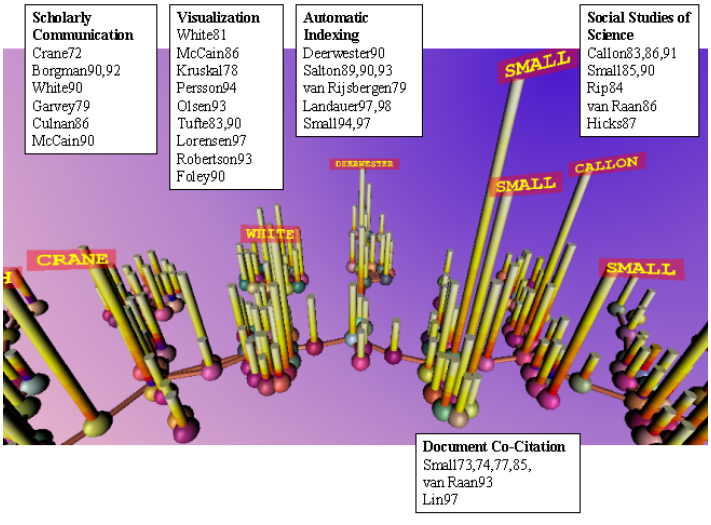
Landscape View of Author Co-citation Map

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.



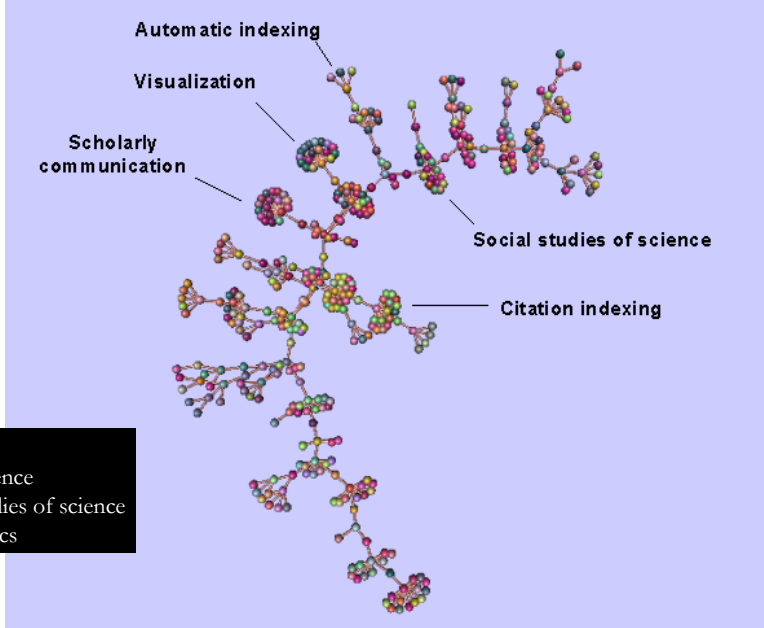
The Document Co-citation Analysis Map

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.



The Document Co-citation Analysis Map

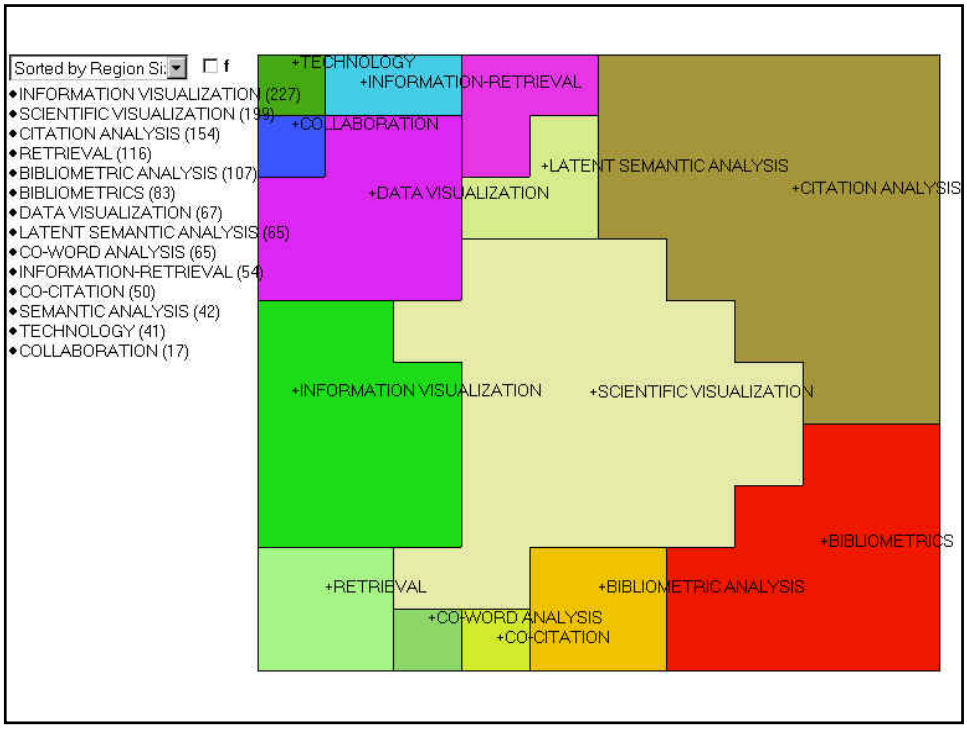
Top-down view. Hand labeling of major clusters.

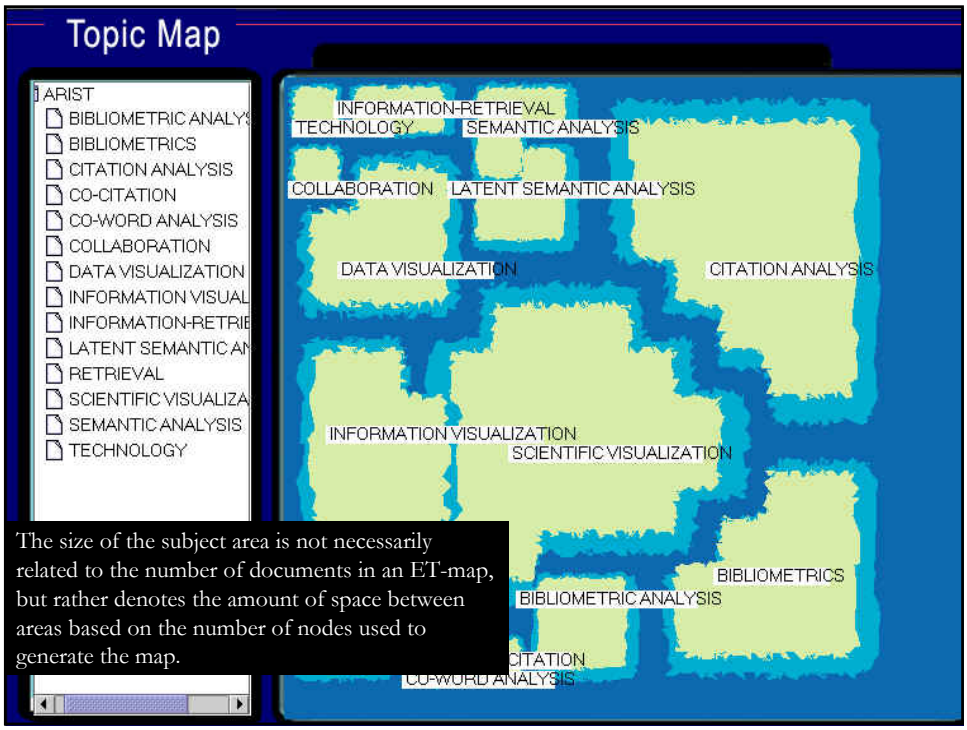


Color code:
red - mapping science
green – social studies of science
Blue – bibliometrics

(2a) ET Map of ARIST Data Set
by Bin Zhu and Hsinchun Chen

- Trained 10x 10 nodes using ID/keyword data of the ARIST data set.
- After training, each node is associated with a list of documents that are semantically similar to each other.
- Each document list is labeled by the most frequently occurring keyword
- Spatial proximity on the map indicates semantic proximity.



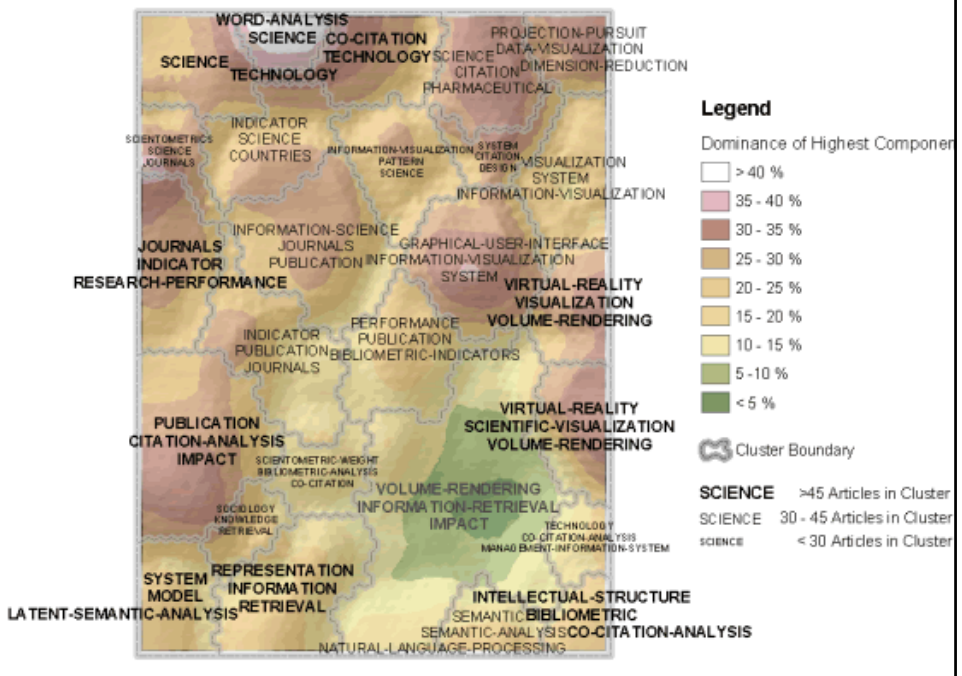


The size of the subject area is not necessarily related to the number of documents in an ET-map, but rather denotes the amount of space between areas based on the number of nodes used to generate the map.

(2b) SOM Map of ARIST Data
by Andre Skupin

- SOM are used to generate domain visualizations in cartographic fashion.
- 40 x 55 node SOM was trained based on ID/keyword list of ARIST data set.
- ArcGIS is used to generate the visualization.
- Dominance of clusters corresponds to number of articles it contains. Higher elevation—i.e., percentage—indicates a very organized, focused, and coherent portion of the information space.
- Labels are automatically assigned based on highly frequent keywords and are drawn within ArcGIS.

Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, Italy, June 16th, 2004.



(3) VxInsight, Sandia National Labs

Next slides show:

- VxInsight citation maps of ARIST data for four different time segments.
- VxInsight co-term and LSA maps of ARIST data.
- VxInsight co-classification map of ARIST data.
- Comparison of maps.

Dot color legend

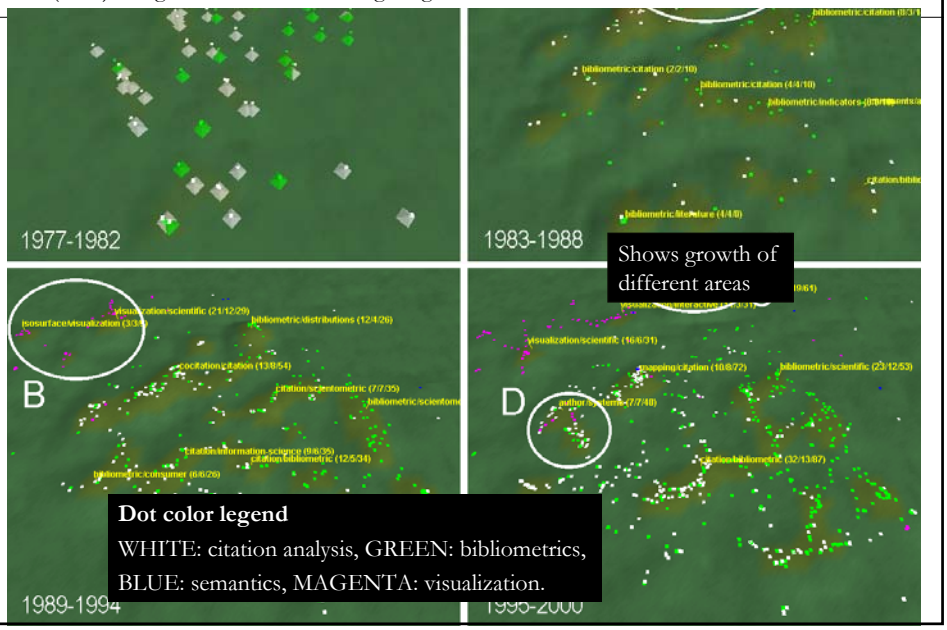
WHITE: citation analysis, GREEN: bibliometrics,
BLUE: semantics, MAGENTA: visualization.

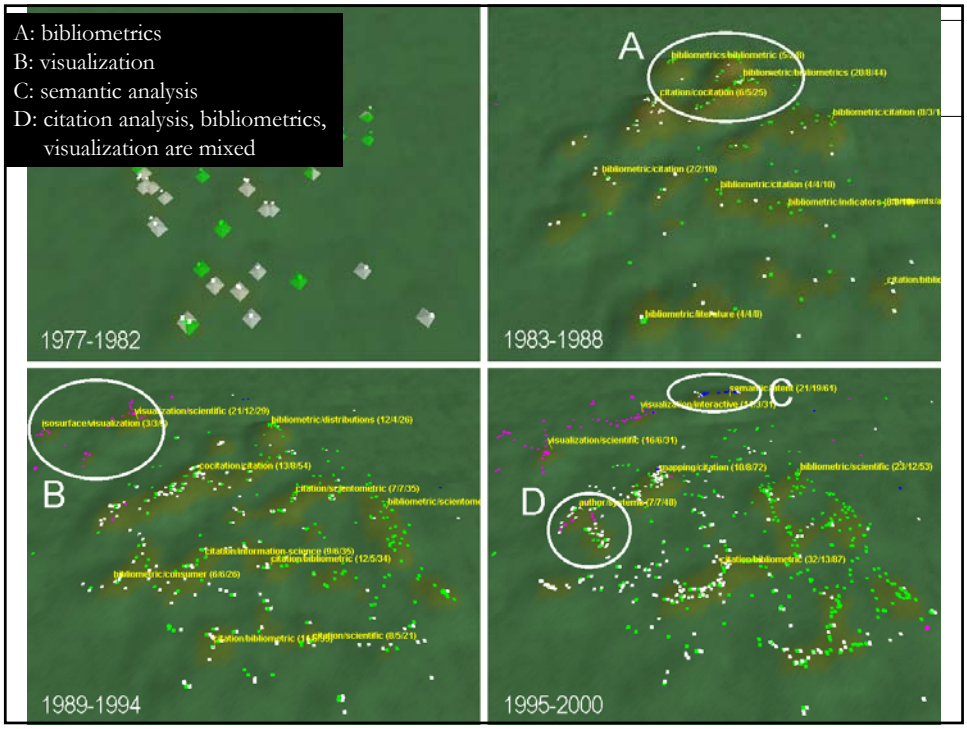
VxInsight Interface

Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, Italy, June 16th, 2004.

1. VxInsight citation maps of ARIST data for four different time segments.

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.

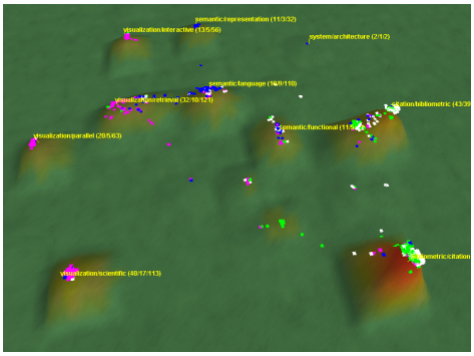


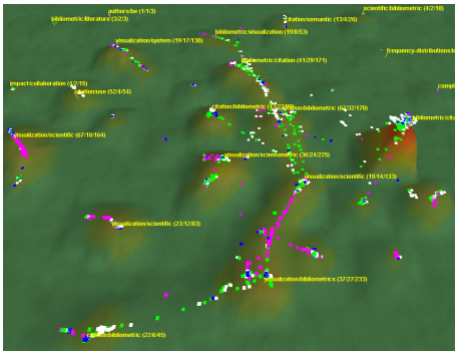


2. VxInsight co-term and LSA maps of ARIST data

Co-term map is based on a cosine similarity using ISI keywords.

LSA was applied over title words to generate a document-by-document similarity matrix. Only similarity values ≥ 0.9 were used in VxOrd FDP to generate the map.



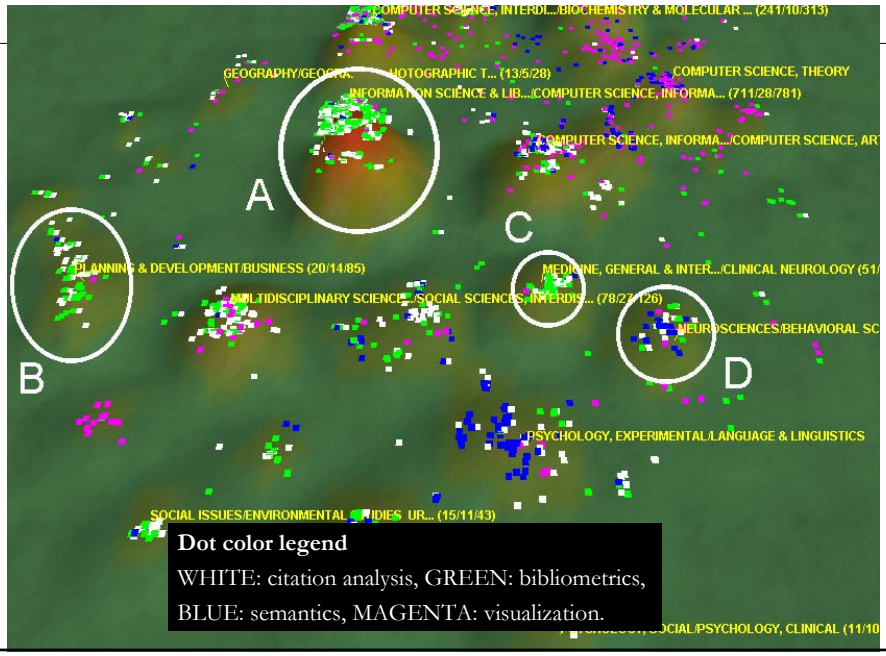


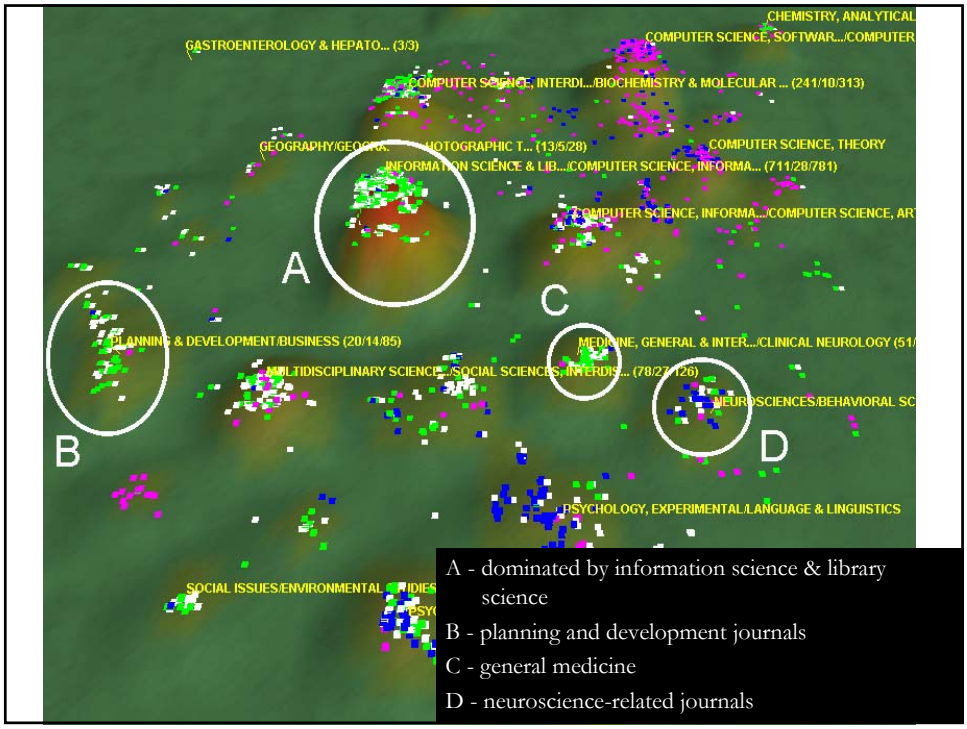
Dot color legend
WHITE: citation analysis, GREEN: bibliometrics,
BLUE: semantics, MAGENTA: visualization.

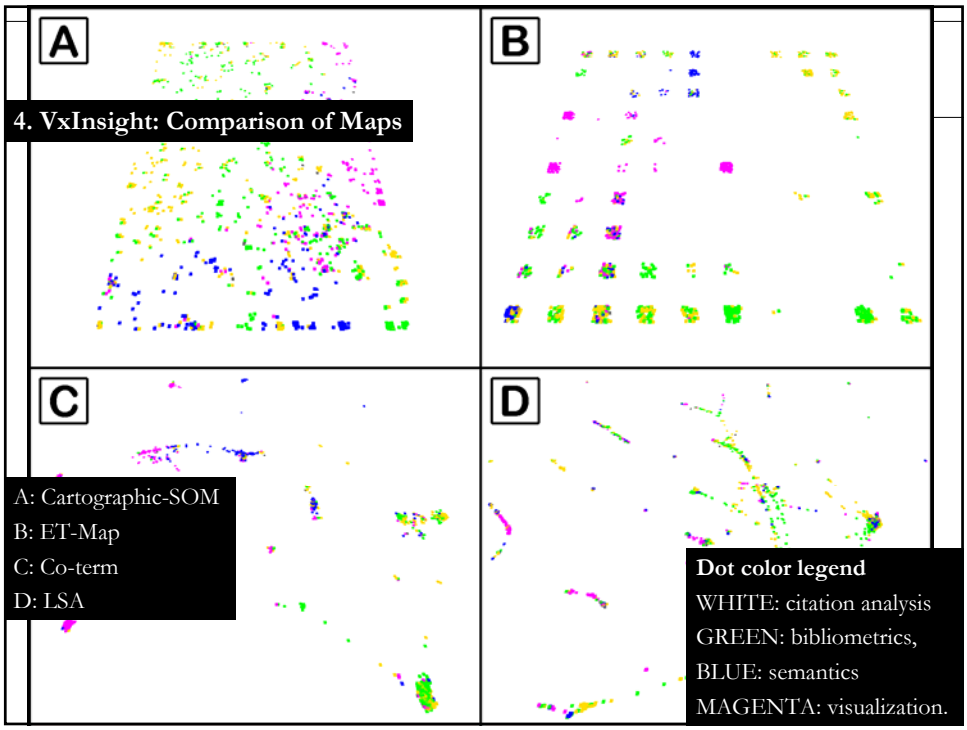
Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, Italy, June 16th, 2004.

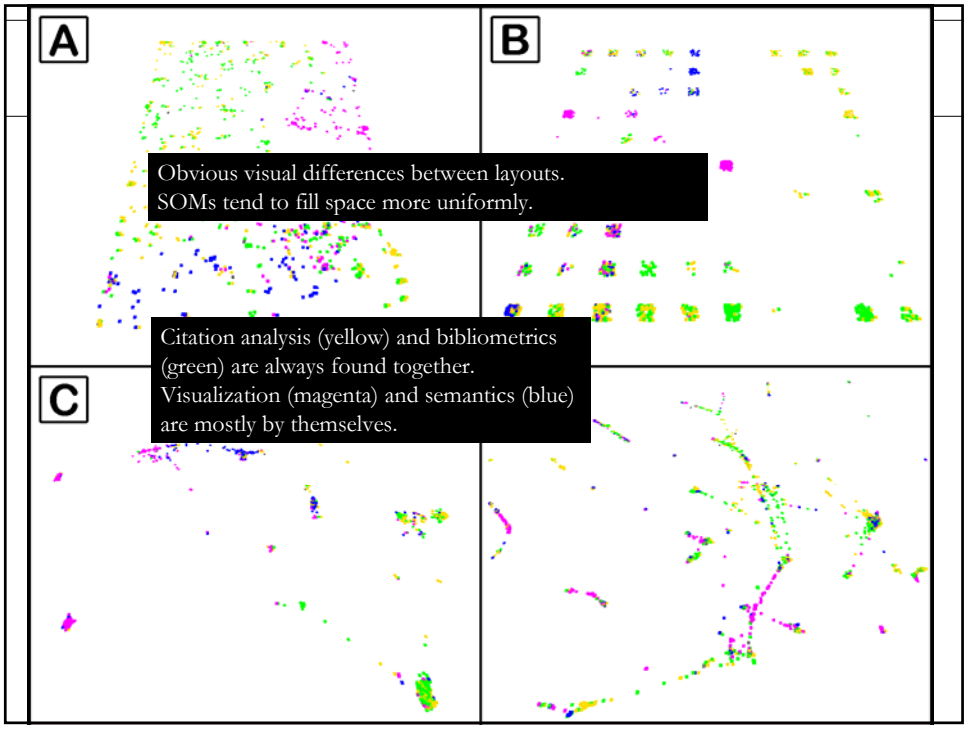
3. VxInsight co-classification map of ARIST data

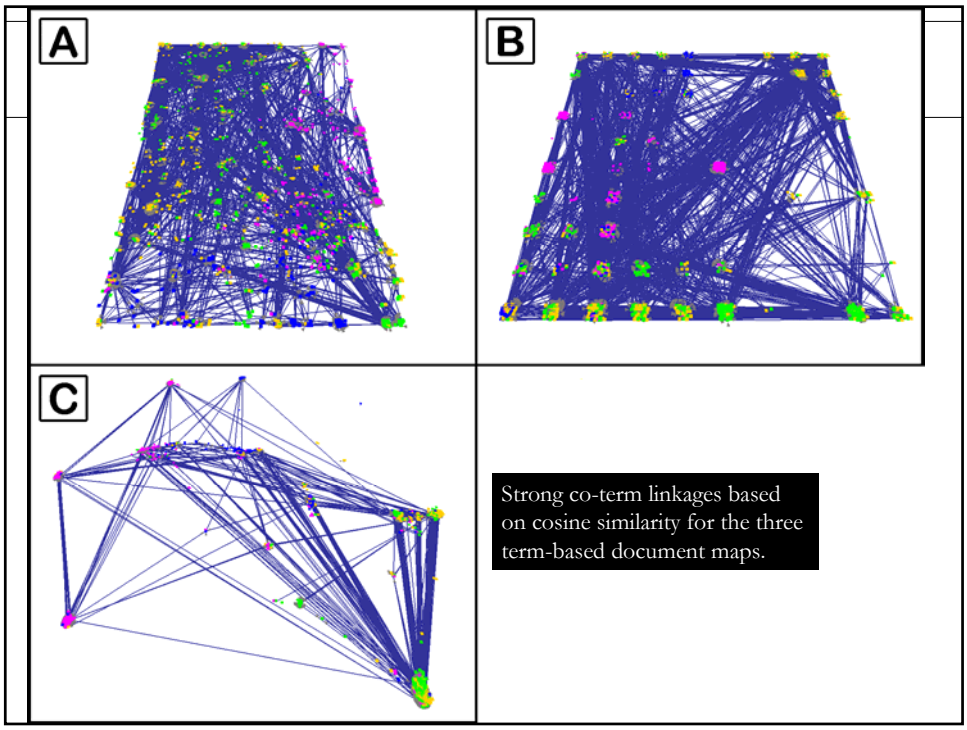
based a cosine similarity from the ISI journal classifications for each article.




CHEMISTRY, ANALYTICAL
COMPUTER SCIENCE, SOFTWAR.../COMPUTER
GASTROENTEROLOGY & HEPATO... (3/3)
COMPUTER SCIENCE, INTERDL.../BIOCHEMISTRY & MOLECULAR ... (241/10/313)
COMPUTER SCIENCE, THEORY
INFORMATION SCIENCE & LIB.../COMPUTER SCIENCE, INFORMA... (711/28/781)
COMPUTER SCIENCE, INFORMA.../COMPUTER SCIENCE, AR
A
C
B
D
PLANNING & DEVELOPMENT/BUSINESS (20/14/85)
MEDICINE, GENERAL & INTER.../CLINICAL NEUROLOGY (51/
MULTIDISCIPLINARY SCIENCE
SOCIAL SCIENCES, INTERDIS... (78/27/126)
NEUROSCIENCES/BEHAVIORAL SC
PSYCHOLOGY, EXPERIMENTAL/LANGUAGE & LINGUISTICS
SOCIAL ISSUES/ENVIRONMENTAL
A - dominated by information science & library science
B - planning and development journals
C - general medicine
D - neuroscience-related journals


A
B
C
D
4. VxInsight: Comparison of Maps
A: Cartographic-SOM
B: ET-Map
C: Co-term
D: LSA
Dot color legend
WHITE: citation analysis
GREEN: bibliometrics,
BLUE: semantics
MAGENTA: visualization.


A
B
C
Obvious visual differences between layouts.
SOMs tend to fill space more uniformly.
Citation analysis (yellow) and bibliometrics (green) are always found together.
Visualization (magenta) and semantics (blue) are mostly by themselves.


A
B
C
Strong co-term linkages based on cosine similarity for the three term-based document maps.

Maps show

- How research on *Visualizing Knowledge Domains* grows out of *semantic analysis/indexing/mapping*, *citation analysis*, *bibliometrics*, and *visualization*.
- That there is interaction between the groups of researchers and their literature in citation analysis and bibliometrics while visualization and semantics are mostly by themselves.

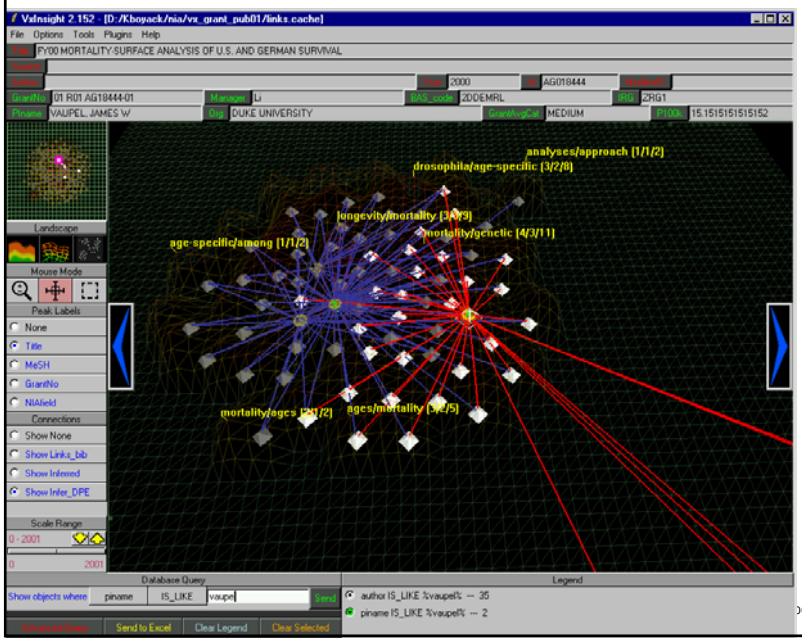
Summary

- Analysis and layout comparisons do not show that any one type of similarity method and layout are better than others for producing domain visualizations.
- Trade-offs are involved and researcher should use the combination of similarity and layout techniques that are likely to aid in answering the questions at hand.
- We encourage researchers and practitioners 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.

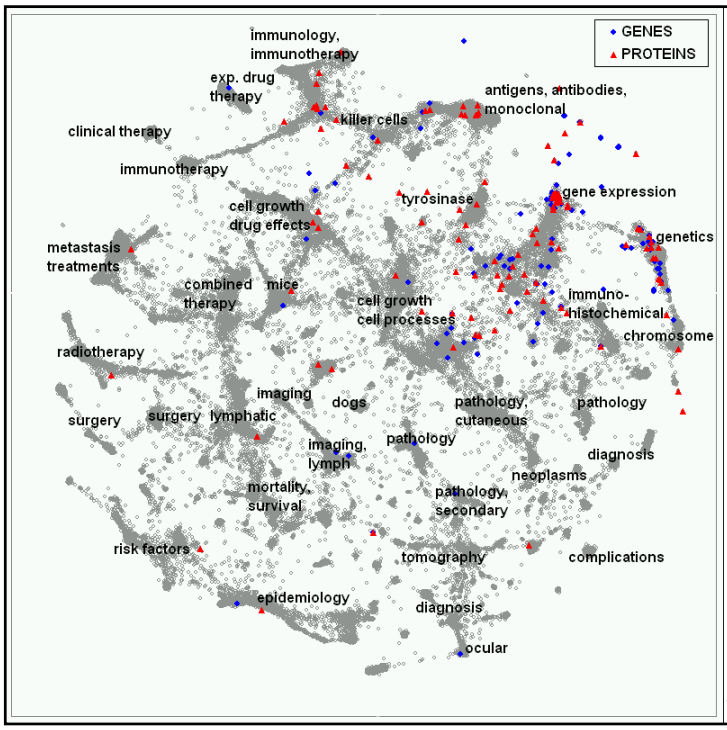
Research Projects

Indicator-Assisted Evaluation and Funding of Research

Visualizing the influence of grants on the number and citation counts of research papers (Boyack & Börner, 2003)



004.



Mapping Medline Papers, Genes, and Proteins Related to Melanoma Research

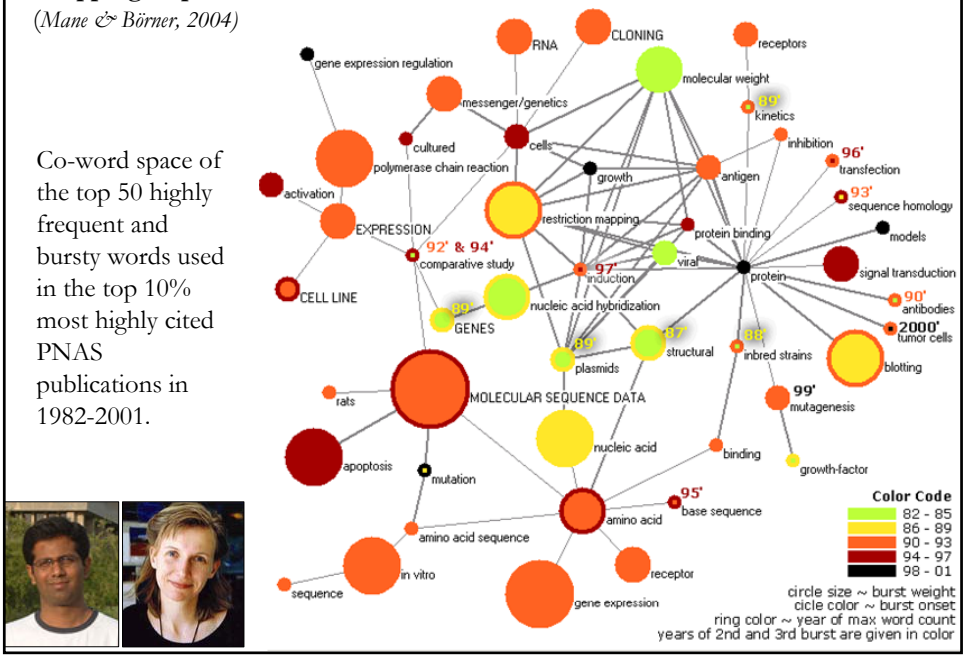
(Boyack, Mane & Börner, 2004)

June 16th, 2004.

Mapping Topic Bursts

(Mane & Börner, 2004)

Co-word space of the top 50 highly frequent and bursty words used in the top 10% most highly cited PNAS publications in 1982-2001.



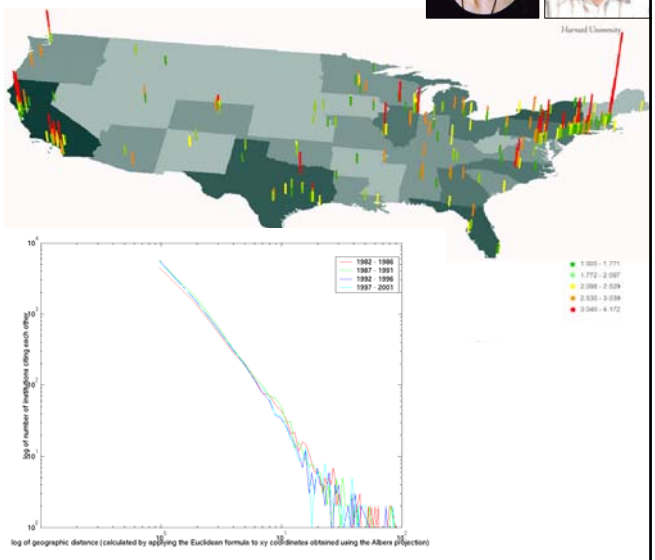
Information Diffusion Patterns

(Börner & Penumarthy, 2004)

Top 500 most highly cited U.S. institutions.

Each institution is assumed to produce and consume information.

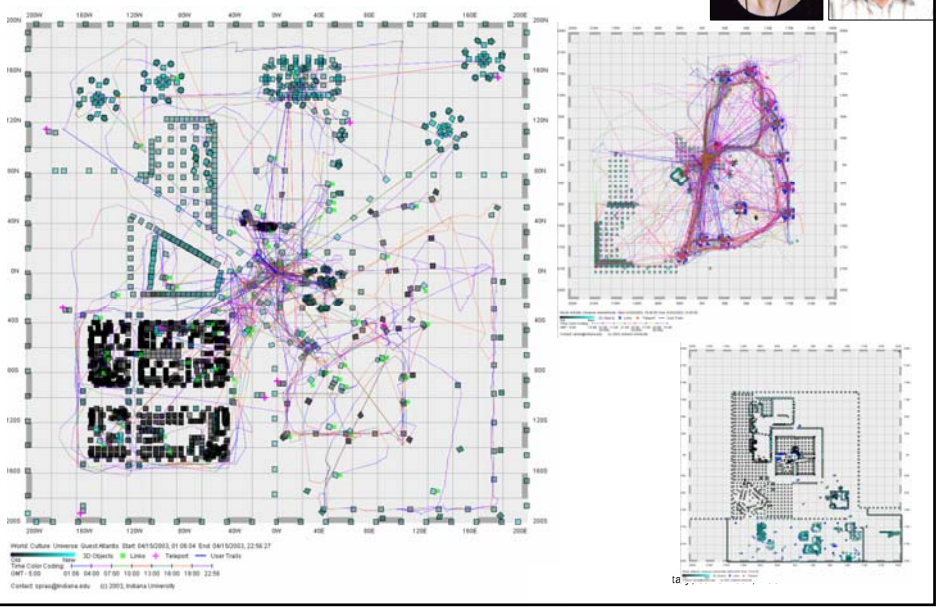
Does Internet lead to more global citation patterns, i.e., more citation links between papers produced at geographically distant research instructions?



Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, Italy, June 16th, 2004.

Visualizing Social Diffusion Patterns in 3D (Virtual Worlds)

(Börner & Penumarthy, 2003 & 2004)



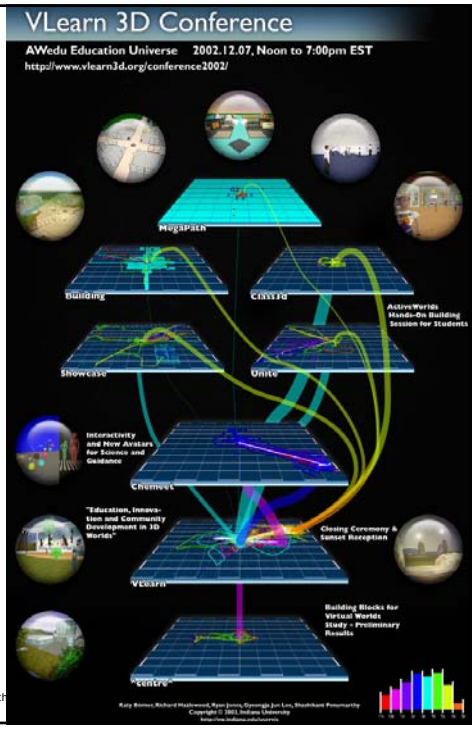
VLearn 3D Vis

(Börner, Hazlewod, Jones, Lee & Penumarthy, 2003)

Temporal-spatial distribution of Conference attendees

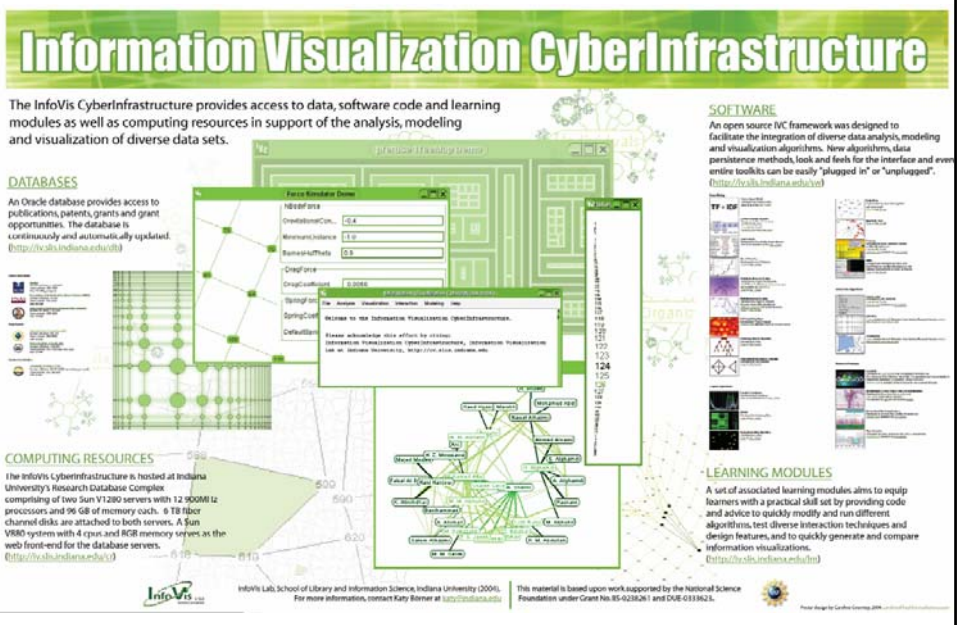
- Conference worlds are represented by square, perspective maps, each labeled by its name.
- Worlds accessed at the beginning of the conference are placed at the bottom, worlds accessed later toward the top.
- Next to each world is a circular snapshot of the virtual venue. Short descriptions of the main sessions are added as text.
- Major jumps between worlds are visualized by transparent lines. The thickness of each line corresponds to the number of traveling users. Color coding was used to denote the chronological paths of the conference sessions.

Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for th

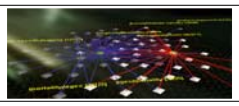


Infrastructure

Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, Italy, June 16th, 2004.



Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, Italy, June 16th, 2004.



InfoVis CyberInfrastructure

InfoVis CyberInfrastructure

A Data-Code-Compute Resource for Research and Education in Information Visualization

Home | Learning Modules | Software | Data Bases | Compute Resources | References

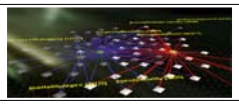
This web site provides access to different software packages easing the exploration, modification, comparison, and extension of data mining and information visualization algorithms. Diverse software packages were bundled into learning modules. Links to diverse databases, compute resources, and references are provided as well. It is our hope that the community will adopt this resource to foster Information Visualization education and research. This site is work in progress. A very first version was released at IEEE InfoVis in October 2003.

Acknowledgements

The Information Visualization Software Repository was created in 2000 and has since then been used to teach the Information Visualization class at Indiana University. Katy Börner, Yuezheng Zhou, and Jason Baumgartner implemented the very first algorithms. In Summer 2003, Jason Baumgartner, Nihar Sheth, and Nathan J. Deckard lead a project to design a XML toolkit that enables the serialization and parallelization of commonly used data analysis and visualization algorithms. Contributions of software packages and implementation work are acknowledged on the respective software pages.

Support comes from the School of Library and Information Science, Indiana University's High Performance Network Applications Program, an Academic Equipment Grant by SUN Microsystems, SBC (formerly Ameritech) Fellow Grant, and the National Science Foundation under DUE-0333623 and IIS-0238261.

Information Visualization Cyberinfrastructure @ SLIS, Indiana University
Last Modified October 16, 2003



InfoVis Learning Modules

InfoVis CyberInfrastructure

A Data-Code-Compute Resource for Research and Education in Information Visualization

Home | Learning Modules | Software | Data Bases | Compute Resources | References

This web site provides access to different software packages easing the exploration, modification, comparison, and extension of data mining and information visualization algorithms. Diverse software packages were bundled into learning modules. Links to diverse databases, compute resources, and references are provided as well. It is our hope that the community will adopt this resource to foster Information Visualization education and research. This site is work in progress. A very first version was released at IEEE InfoVis in October 2003.

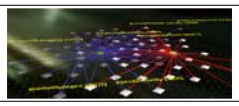
Acknowledgements

The Information Visualization Software Repository was created in 2000 and has since then been used to teach the Information Visualization class at Indiana University. Katy Börner, Yuezheng Zhou, and Jason Baumgartner implemented the very first algorithms. In Summer 2003, Jason Baumgartner, Nihar Sheth, and Nathan J. Deckard lead a project to design a XML toolkit that enables the serialization and parallelization of commonly used data analysis and visualization algorithms. Contributions of software packages and implementation work are acknowledged on the respective software pages.

Support comes from the School of Library and Information Science, Indiana University's High Performance Network Applications Program, an Academic Equipment Grant by SUN Microsystems, SBC (formerly Ameritech) Fellow Grant, and the National Science Foundation under DUE-0333623 and IIS-0238261.

Information Visualization Cyberinfrastructure @ SLIS, Indiana University
Last Modified October 16, 2003

Katy Börner: Mapping Interdisciplinary Research Domains. Parmenides Center for the Study of Thinking, Island of Elba, Italy, June 16th, 2004.



InfoVis Learning Modules: Design

InfoVis CyberInfrastructure

A Data-Code-Compute Resource for Research and Education in Information Visualization

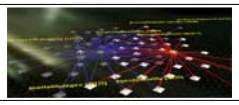
Home | Learning Modules | Software | Data Bases | Compute Resources | References

Learning Modules

Most information visualizations are highly interactive. While a number of excellent textbooks exist, the two-dimensional printouts on paper often cannot convey their true visual appearance and interactive performance. Several textbooks come with accompanying web sites that contain snapshots of user interfaces as well as animations and movies. However, none of them facilitates the exploration, application, evaluation, and comparison of algorithms.

This web page will provide access to a number of learning modules. Each learning module comes with an:

- Introduction
- Discussion of Existing Algorithms
- Learning Task
  - A challenging scenario to use the code in the XML Toolkit to visualize a data set to support a specfic user group.
- Programming Exercise
  - An programming exercise plus an explanation of possible solutions.
- Opportunities & Challenges, and
- References to research papers, online demos, (commercial) applications)



Visualizing Tree Data

http://iv.slis.indiana.edu/lm/lm-trees.html

InfoVis CyberInfrastructure

A Data-Code-Compute Resource for Research and Education in Information Visualization

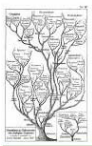
Home | Learning Modules | Software | Data Bases | Compute Resources | References

Learning Modules > Visualizing Tree Data

Description | Usage Hints | Learning Task | Discussion | References | Acknowledgments

Description

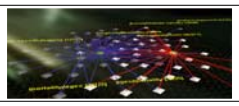
Many data sets come in tree format. There are family trees, organizational charts, classification hierarchies, and directory structures. The figure below shows an inheritance tree by Ernst Haeckel ('Stammbaum' in German). Read also To Draw a Tree by Pat Hanrahan.



Click image for larger version

A tree graph is a set of straight line segments (edges) connected at their ends containing no closed loops (cycles). You can also call it a simple, undirected, connected, acyclic graph (or, equivalently, a connected forest). A tree with n nodes has n-1 graph edges. All trees are bipartite graphs.

Many trees have a root node and are called rooted trees. Trees without a root node are called free trees. Subsequently, we will only consider rooted trees. In rootes trees, all nodes except the root node have only one parent node. Nodes which have no children are called leave nodes. All other nodes are referred to as intermediate nodes.



Student's Project Results

User & Task Analysis for Visualizing Tree Data

- Visualizing the structure of IU's Decision Support System
- Visualizing the co-occurences of keywords in DLib Magazine articles.
- Visualization of the Java API
- Visualizing the the Library of Cor
  retrieve legal materials in a library.

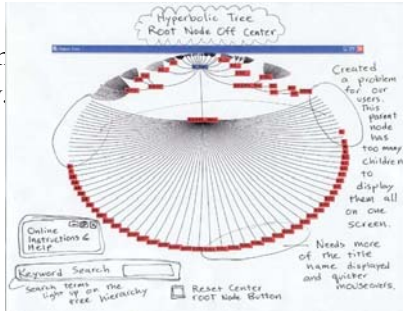
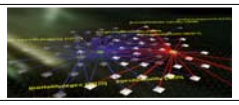


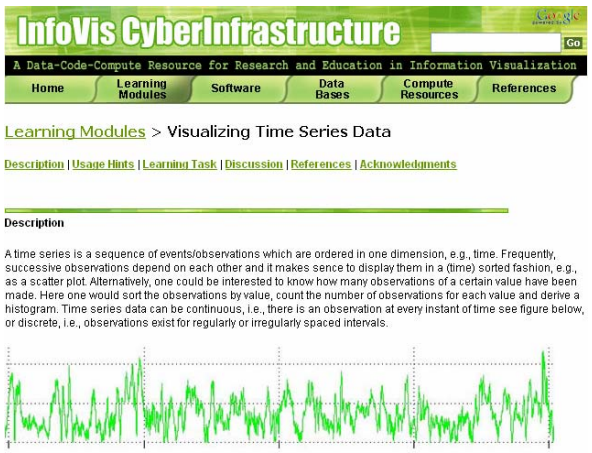
Image by Peter Hook and Rongke Gao

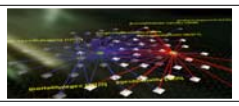
See Handin pages at



Time Series Analysis & Visualization

http://iv.slis.indiana.edu/lm/lm-time-series.html





Student's Project Results

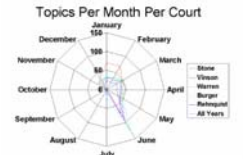
Time Series Analysis & Visualization

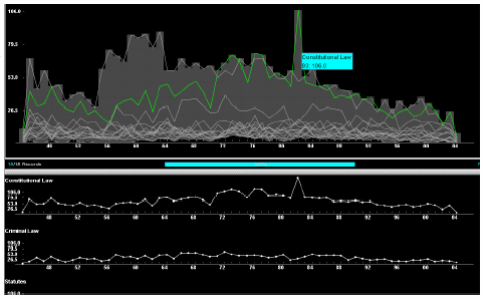
- Using Timesearcher and the Burst Detection Algorithm to Analyze the Stock Market from 1925 to 1945
- Applying Burst and TimeSearcher to Chat Data
- Lab Access Trends
- Quest Atlantis Chat Log Data

See Handin pages at
http://ella.slis.indiana.edu/~katy/handin/L579-S04/cgi/handinlogin.cgi

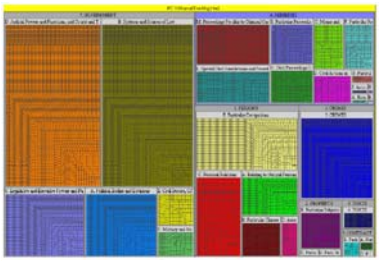
Visualizing the Work of the United States Supreme Court Based on Time Data and Top Level West Topics

by Peter A. Hook & Rongke Gao

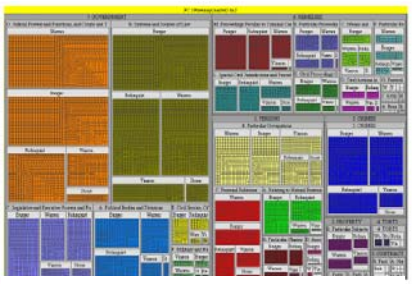




Top fifteen most occurring topics from 1944 to 2004 in Timesearcher



All topics grouped by West Category and Sub-Category grouped over the entire lengths of the data set



All topics by West Category and Sub-Category grouped corresponding to the five chief justices

Visualizing Niches of the Blog Universe

BY Mike Tyworth and Elijah Wright