Visualizing a Knowledge Domain's Intellectual Structure

To make knowledge visualizations clear and easy to interpret, we have developed a method that extends and transforms traditional author co-citation analysis by extracting structural patterns from the scientific literature and representing them in a 3D knowledge landscape.

Chaomei Chen
Ray J. Paul
Brunel University

Visualizing the entire body of scientific knowledge and tracking the latest developments in science and technology have intrigued generations of scientists, philosophers, government officials, librarians, and publishers. Advances in information visualization offer promising tools for presenting knowledge structures and their development in an increasingly intuitive way.[1]

The scientific literature provides ingredients ripe for knowledge visualization. Researchers commonly focus on significant structural patterns in knowledge discovery, information retrieval, and other disciplines that may offer insights into the nature of underlying interrelationships such as those among authors of scholarly publications,[2] documents,[3] and journals.[4]

We describe an approach to visualizing a knowledge domain's intellectual structures that extracts structural patterns from the scientific literature and represents them in a 3D knowledge landscape. This approach extends and transforms traditional author co-citation analysis (ACA) into a knowledge-visualization and domain-analysis tool.

KNOWLEDGE VISUALIZATION

Several existing systems address common knowledge-visualization issues, such as selecting appropriate similarity metrics and displaying high-dimensional structures. SemNet, introduced in the 1980s, produces 3D graphic representations of large knowledge bases to help users grasp complex relationships.[5] SemNet's design focuses on the graphical representations of three types of components:

- the identity of individual elements in a large knowledge base,
- the relative position of an element within a network context, and
- explicit relationships between elements.

SemNet represents elements of a Prolog rules knowledge base as labeled rectangles connected by lines or color-coded arcs. A Prolog module, which contains a subset of the logic programming language's rules, thus appears as a rectangle labeled with the module's name. In SemNet, the closeness between two rectangles indicates the strength of the connection between their modules. To show how the knowledge base works, SemNet's designers experimented with various techniques such as multidimensional scaling (MDS), simulated annealing, fisheye views, and even a sprite that travels down arcs between rectangles.

Visualization tools

The Spatial Paradigm for Information Retrieval and Exploration, developed by Pacific Northwest National Laboratory, provides a classic example of information visualization.[3] Spire consists of a suite of visualization tools for browsing large sets of documents—also called a document collection or corpus—and includes a well-known visualization view called *Theme-scape*. This tool creates an abstract, 3D landscape view of a document corpus. A thematic terrain simultaneously communicates both the primary themes of the underlying document collection and a measure of their relative prevalence in the collection. Thematic peaks and valleys in Themescape produce a simplified representation of a document corpus's complex content.

VxInsight,[6] a knowledge visualization tool developed by Sandia National Laboratories, groups data elements from very large data sets by similarity. The

0018-9162/01/$10.00 © 2001 IEEE

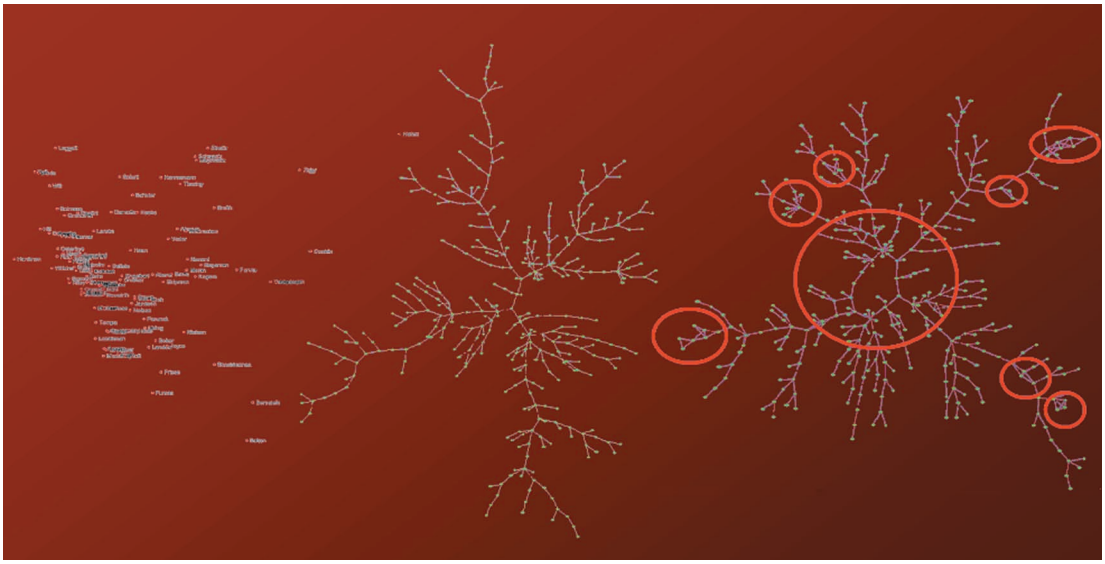

Figure 1. The ACM hypertext citation network of 367 authors and 61,175 links as it appears in the MDS, MST, and Pathfinder solutions (from left to right), which consist of 0, 366, and 398 links, respectively. Circles in Pathfinder highlight cyclic links.

tool uses the height of a mountain in a 3D virtual landscape to portray the density of data elements distributed beneath. Using a subset of the Science Citation Index (SCI), a citation database from the Institute for Scientific Information (ISI), researchers have applied VxInsight to the visualization of nuclear physics. The tool leverages similarities between two documents, which are proportional to the extent that the documents have common citation links, to generate visualizations using a combination of eigenvector-based and force-directed placement solutions. In addition to SCI, ISI provides other citation databases, including the Social Sciences Citation Index.

Henry Small[4] presents a visualization of the scientific literature based on journal co-citation patterns derived from these ISI citation databases. His work covers the widest range of scientific knowledge domains so far. In addition to these perspectives of knowledge visualization, groupings of scientists by their expertise and significant contributions to a knowledge domain can provide a unique representation of the domain's structure. If the groupings reflect the intellectual connections as perceived by the scientists in the same field, visualizing such intellectual structures will likely reveal valuable insights into the knowledge structure.

We have developed a generic knowledge visualization approach that extends the traditional ACA approach and enhances it with a 3D knowledge landscape.

Author co-citation patterns

ACA, a special type of citation analysis, focuses on intellectual connections between authors as reflected through the scientific literature. The author co-citation relationship links two authors by how often other authors reference their work together. Author co-citation patterns provide the basis for constructing an alternative view to a knowledge structure.

Research into ACA has demonstrated its potential as a powerful tool for visualizing the intellectual structures of specific disciplines. ACA focuses on how individual authors perceive the intellectual structure of their own subject domain. In an in-depth author co-citation study, Howard White and Katherine McCain[2] analyzed the information-science domain based on author co-citation data drawn from 12 key journals in the field. Their work represents the state of the art in ACA, which typically uses factor analysis and clustering techniques to determine intellectual groupings, then depicts the results as MDS solutions.

We have developed a generic approach that extends traditional ACA analysis by integrating structural modeling and information visualization techniques to provide a 3D knowledge landscape based on citation patterns. In particular, we introduce the following steps to extend conventional ACA to visualize intellectual structures:

- replace MDS with the Pathfinder network scaling technique to display interrelationships and local structures explicitly and more accurately,
- visualize the intellectual groupings determined by factor analysis in traditional ACA, and
- evaluate the citation impact in the context of a co-citation network.

Pathfinder network scaling

Pathfinder network scaling is a structural-modeling tool developed by cognitive psychologists.[7] It provides an effective way to extract the most essential relationships from a given set of proximity data and simplifies a network by retaining only the strongest paths. Pathfinder uses a filtering criterion known as the *triangle inequality condition* to determine whether to remove or retain each link in the original network.[8] Triangle inequality requires that the length of a path connecting two points in the network should not be longer than the length of other alternative paths connecting the two points, but go through extra intermediate points.

Pathfinder offers a better alternative to traditional layout and link-reduction methods such as MDS and minimum spanning trees (MSTs). MDS provides no

explicit representations of links, making it difficult to interpret the nature of each dimension in an MDS solution. Pathfinder explicitly represents the strongest links, and the interpretation relies on linkage instead of relative positions along each dimension. Scientists commonly visualize information as a general network in the form of an MST. Typically, a Pathfinder network forms a superset of an MST that contains links from all possible MSTs. This feature maintains the underlying structure's semantic integrity.

Figure 1 illustrates the differences between MDS, MST, and Pathfinder solutions generated for the same set of citation data: the ACM Hypertext Conference Series proceedings from 1989 to 1999. Our previous work provides a detailed description of the ACM hypertext analysis.[9]

FOUR-STEP PROCEDURE

Traditional ACA uses factor analysis to identify intellectual groupings—known as specialties—from author co-citation data, then it uses MDS to depict such groupings. We enhance and extend this procedure by introducing Pathfinder network scaling to replace MDS. We also integrate Pathfinder and factor analysis to visualize specialties in the underlying knowledge domain. Further, we visualize the citation frequency of each leading scientist in the context of specialties so that we can track changes to that author's influence over time.

Figure 2 shows our approach's four-step procedure. First, we select authors whose work has received citations above a predetermined threshold. The intellectual groupings of these authors provide snapshots of the underlying knowledge domain. We compute 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[2] believe that such correlation coefficients best capture an author's citation profile.

Second, we apply Pathfinder network scaling to the network that the correlation matrix defines. Although factor analysis is a standard ACA practice, MDS and factor analysis rarely appear in the same graphical representations in traditional author co-citation analysis. To make knowledge visualizations clear and easy to interpret, we 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, which forms the third step.

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. Figure 3 shows the construction of a 3D knowledge landscape.

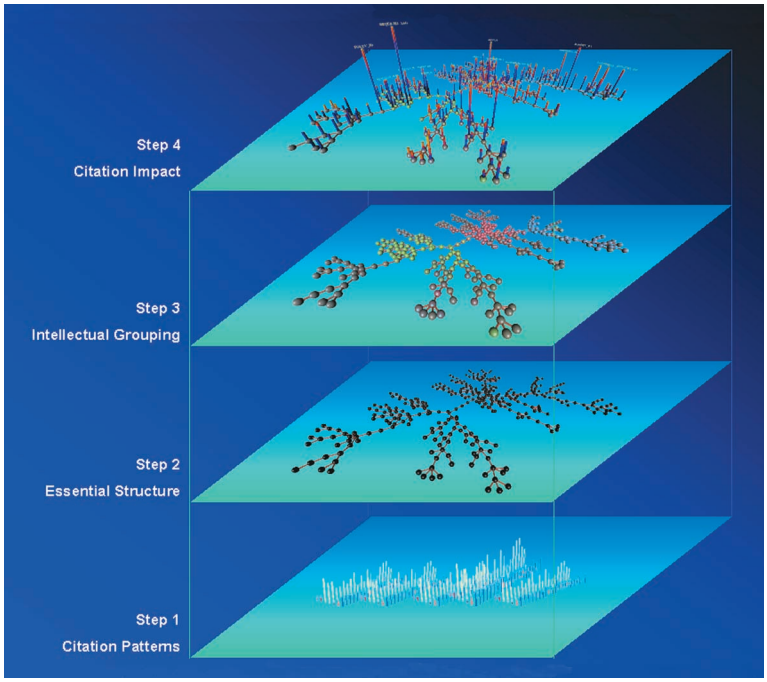

Figure 2. A four-step procedure for visualizing intellectual structures. The process starts from the lowest level and works its way up by incrementally overlaying more visual-spatial features at each consecutive step to clarify the essence of intellectual structures.

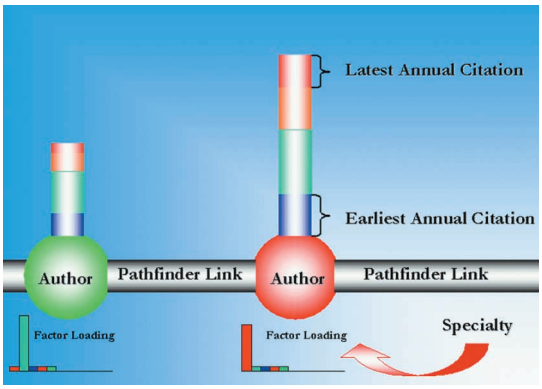

Figure 3. Design of a citation and co-citation landscape. The three-dimensional model consists of an author co-citation network on a two-dimensional plane plus the historical citation profile for each author in the third dimension.

COMPUTER GRAPHICS VISUALIZATION

A computer graphics visualization project demonstrates our four-step procedure in action.

We began by studying author co-citation patterns found in *IEEE Computer Graphics and Applications* magazine for a period of 18 years. The SCI citation database covers *CG&A* from Volume 2 in 1982 to Volume 19 in 1999.

The *CG&A* citation data include articles written by 1,820 authors and co-authors. Counting the first author only, these *CG&A* authors cited a total of 10,292 unique articles written by 5,312 authors. Among them, we entered into the author co-citation analysis only the 353 authors who received more than five citations in *CG&A*. Although this snapshot derives from a limited viewpoint—the literature of

Table 1. Top 10 leading scientists in the five largest intellectual groups.

Authors	F1	Authors	F2	Authors	F3	Authors	F4	Authors	F5
T Whitted	0.895	RB Tilove	0.876	MA Sabin	0.819	D Gordon	0.663	NI Badler	0.635
L Williams	0.890	MA Wesley	0.873	W Boehm	0.776	G Frieder	0.651	D Zeltzer	0.600
ME Lee	0.886	HB Voelcker	0.866	RH Bartels	0.774	JK Udupa	0.649	B Mandelbrot	0.597
DS Kay	0.885	CM Brown	0.861	PE Bezier	0.774	LT Cook	0.649	CW Reynolds	0.578
NL Max	0.878	JW Boyse	0.859	C deBoor	0.770	E Artzy	0.628	FI Parke	0.577
RL Cook	0.875	TC Woo	0.857	SA Coons	0.765	LS Chen	0.627	D Gordon	0.570
DR Warn	0.874	M Mantyla	0.851	TNT Goodman	0.759	SM Goldwasser	0.627	WW Armstrong	0.572
JC Hourcade	0.874	YT Lee	0.845	RF Riesenfeld	0.754	RA Reynolds	0.623	D Thalmann	0.564
BT Phong	0.862	G Markowsky	0.834	B Joe	0.754	H Fuchs	0.622	J Wilhelm	0.557
RA Hall	0.858	AP Morgan	0.829	IJ Schoenberg	0.747	P Dev	0.601	JK Udupa	0.556

computer graphics certainly stretches beyond the scope of *CG&A*— intellectual groupings of these 353 authors provide the basis for visualizing the computer graphics knowledge domain.

The original author co-citation network contains as many as 28,638 links, which constitutes 46 percent of all possible links, excluding self-citations. Because this many links would clutter visualizations, we applied Pathfinder network scaling to reduce their number to 355.

Author co-citation structures

We used a 3D virtual landscape to represent author co-citation structures. The most influential scientists in the knowledge domain appear near the intellectual structure's center. In contrast, researchers who have unique expertise tend to appear in peripheral areas. The virtual landscape also lets users access further details regarding a particular author in the intellectual structure, such as a list of the author's most-cited work, abstracts, and even the full content of that author's articles.

We enhanced the network by coloring it according to the results generated using principal component analysis (PCA). PCA identified 60 specialties in computer graphics. The largest (rendering and ray tracing) and second-largest (computer vision) accounted for 13 percent and 11 percent of the variance, respectively. The five largest specialties accounted for 39 percent of the variance. Remaining specialties are relatively small. Table 1 lists the names of the 10 most predominant members in each of the five largest intellectual groups.

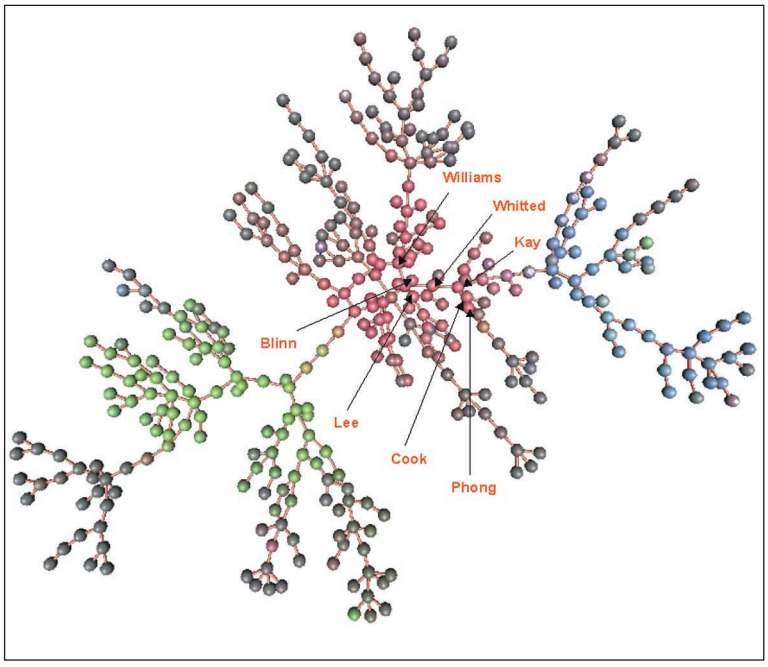

Figure 4. Leading members in rendering and ray tracing, the largest specialty.

ResearchIndex citations

To determine the nature of the predominant intellectual groups, we used the ResearchIndex citation system—formerly known as CiteSeer—developed at the NEC Research Institute[10] to examine the citation context of leading authors in these specialties. For example, Turner Whitted, a leading member of the rendering and ray-tracing group, the largest specialty, published his pioneering ray-tracing article in 1980 in *Communications of the ACM*. Subsequently, authors of *CG&A* cited this article frequently. ResearchIndex reveals that this article appears on more than 50 Web sites and serves as a classic citation of research in ray tracing. If other leading members of the same specialty have published papers on similar topics, we can conjecture that this specialty focuses on ray tracing. Automatic identification of a specialty's nature remains a challenge.

In Figure 4, color coding identifies the specialties by their PCA factor-loading, with the rendering and ray-tracing group appearing in red, computer vision in green, and geometric modeling and computer-aided design in blue. Many smaller specialties reside along the network's rim.

Factor 1: Rendering and ray tracing. Figure 4 identifies seven leading members in the rendering and ray-tracing specialty. Whitted's illumination model for ray tracing, Lance Williams's classification of level of details, the Cook-Torrance lighting model, and the famous Phong shading model provide notable examples of this specialty. Jim Blinn, one of the most cited authors in *CG&A*, also appears in the group, largely due to his popular blobby model for implicit surface modeling.

Factor 2: Computer vision. The second-largest specialty, computer vision, includes experts such as Robert Tilove, Herbert Voelcker, Christopher Brown, and John Boyse. ResearchIndex recorded 263 cita-

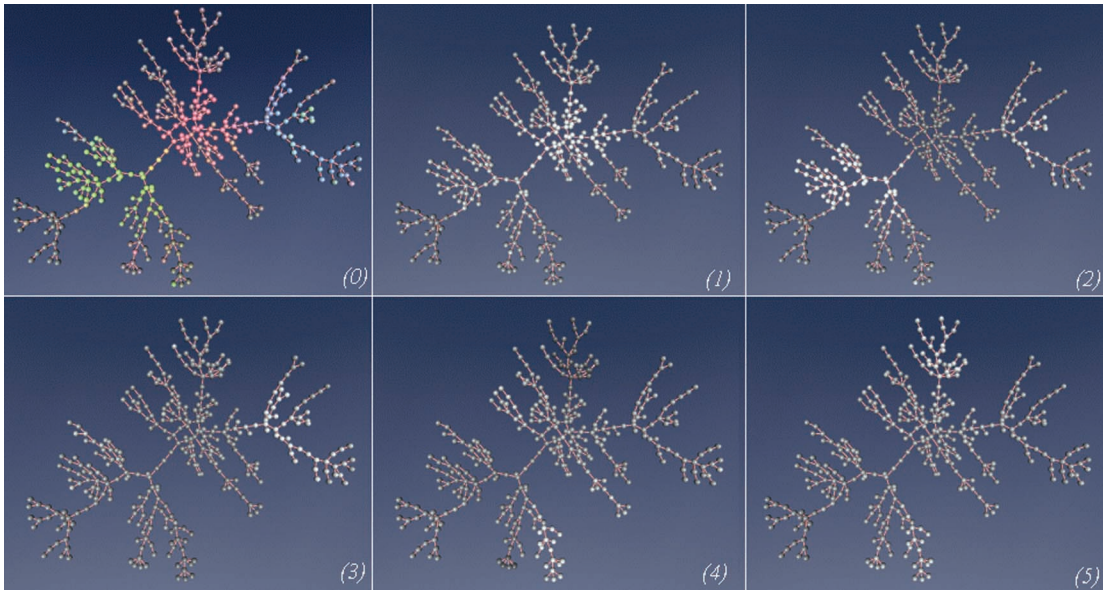

Figure 5. Delineating a high-dimensional intellectual structure. Specialties glow one by one.

tions of a popular computer vision book written by Dana Ballard and Christopher Brown.

Factor 3: Geometric modeling and computer-aided design. The third specialty includes leading scientists—such as Pierre Bezier, Malcolm Sabin, Carl de Boor, and Richard Bartels—whose work focuses on spline-curve and surface representations defined with piecewise polynomial functions. Several concepts in geometric modeling relate to Bezier, including the famous Bernstein-Bezier patches, Bezier clipping, and Bezier triangles. Sabin first generalized Bezier to triangular B-splines, while de Boor published a popular book on splines in 1978. In 1987, Bartels co-authored a book on splines, which scored 66 ResearchIndex citations. This group also includes leading researchers in computer-aided design and geometric modeling.

Factor 4: Volume rendering. Ray-tracing and direct-projection methods provide foundational concepts for directly rendered images. In ray tracing, the light source sends viewing rays through each pixel and integrates the rays throughout the volume. In direct projection, the system projects each cell of the volume onto the screen. The volume-rendering specialty includes researchers such as Dan Gordon and Gideon Frieder. This group also includes image-processing experts such as Jayaram Udupa, whose work focuses on brain image segmentation, and Larry Cook, whose specialty is contour interpolation.

Factor 5: Modeling nature. This fifth-largest specialty includes work in areas that model natural phenomena, such as Norman Badler's simulation of realistic human-figure movements, Benoit Mandelbrot's pioneering fractal geometry, and Craig Reynolds' simulation of birds' flocking behavior. Mandelbrot coined the concept of fractals to describe spiky, irregular, or variegated objects such as coastlines, mountains, and crystals. The computation of a shoreline's length provides a common example of fractal use. The shoreline becomes longer and longer as the map's resolution increases because we must account for every new visible creek at higher resolutions. Frederic Parke also appears in this group for his work on a parameterized facial-modeling approach.

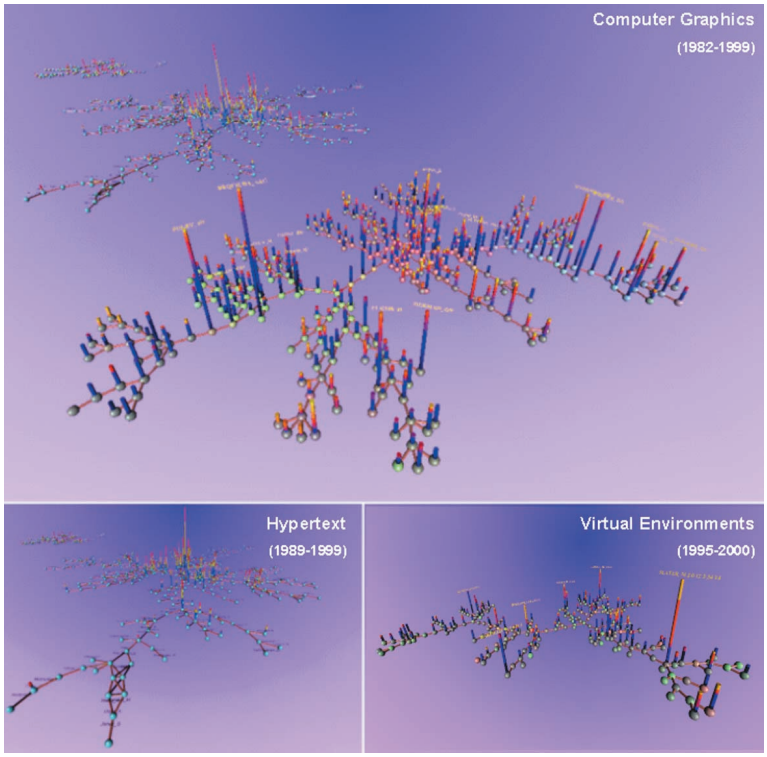

Figure 6. Knowledge landscape of three subject domains: computer graphics, hypertext, and virtual environments.

To delineate the high-dimensional intellectual structure, we animated the glowing of each specialty, one at a time in the network, as Figure 5 shows. The glowing area highlights a particular specialty, which makes it stand out from the others.

Knowledge landscape

Figure 6 shows the intellectual structure as a knowledge landscape. In addition to the specialties derived from the *CG&A* citation data, the virtual landscape also includes hypertext and virtual environments specialties derived from citations in the ACM Hypertext

Conference proceedings from 1989 to 1999, and from *Presence*, a leading scholarly journal in the field of virtual environments, from 1995 to 2000. The 3D landscape invites users to explore rises and falls in citations, the intellectual groupings of authors perceived by thousands of scientists in the same knowledge domain, and the strongest pathways connecting different scientists in the field.

The knowledge landscape visualizes intellectual structures. A virtual landscape like this provides an intuitive gateway for users to access the scientific literature. Researchers new to a field can gain a useful overview by using the knowledge landscape to establish their own mental model of the field and track the development of their own domain.

LESSONS LEARNED

Our approach facilitates visualizing intellectual structures based on widely available sources of citation data. This method augments knowledge-visualization approaches that focus on documents and concepts. The integration of citation and co-citation patterns provides a rich, ecological representation of a knowledge domain. Users can apply such visualizations to discover patterns and make valuable connections between data.

Our experience has also revealed challenges that future work needs to resolve—some common to knowledge visualization in general, others specific to the citation-based approach. For example, reliance on citation data poses a dilemma. On the one hand, citation patterns shed additional insights into a knowledge domain's intellectual structure. On the other hand, citation data constrains the timeliness of visualizing the intellectual structure. Because citation analysis builds on scientists' long-established citation practice, this area will likely pose a long-term challenge. Integrating approaches that focus on Web-based citation resources, such as ResearchIndex,[10] holds promise.

One immediate challenge involves determining how to analyze each specialty's nature effectively, without human intervention. Citation data often lack the information necessary to make such judgments. Few researchers enjoy access to the full content of all articles in electronic format, which provides the needed information. In our example, we retrieved additional information stored in ResearchIndex and examined the context of typical leading-author citations for a given specialty. Future work should investigate this issue in a wider, multidisciplinary context, including digital libraries, natural-language processing, information retrieval, and information visualization.

The use of citation data per se remains controversial. Context-free citation counts cannot replace detailed, context-sensitive citation searches for individual authors. Nevertheless, citation analysis that focuses on prevalent citation patterns can lead to insights. After all, automatically generated intellectual structures are not designed to replace intellectual communications among scientists. Users should avoid overinterpreting intellectual groupings derived from citation patterns alone, no matter how wide the citation data's coverage. Further studies should provide in-depth evaluation of the practical significance of such visualizations.

Our approach to knowledge visualization works particularly well for identifying intellectual groupings based on an extension of the traditional author co-citation analysis. It provides a promising augmentation to existing document- and concept-centered approaches to knowledge visualization. The 3D knowledge landscape of intellectual structures provides users an additional means of exploring and accessing the scientific literature. Our approach has practical implications in multiple disciplines, including knowledge visualization, digital libraries, domain analysis, and particular subject domains.

To grasp the big picture of a knowledge domain, we must consider various factors more thoroughly. We expect that visualizing the intellectual structures of a knowledge domain will provide increasingly powerful tools for tracking the development of scientific knowledge. ✸

Acknowledgments

We thank the reviewers for their helpful comments. This work is supported by research grant GR/L61088 from the Engineering and Physical Sciences Research Council in the UK.

References

1. C. Chen, *Information Visualisation and Virtual Environments*, Springer-Verlag, London, 1999.
2. H.D. White and K.W. McCain, "Visualizing a Discipline: An Author Co-citation Analysis of Information Science," 1972-1995, *J. Am. Soc. Information Science*, vol. 49, no. 4, 1998, pp. 327-356.
3. J.A. Wise et al., "Visualizing the Nonvisual: Spatial Analysis and Interaction with Information from Text Documents," *IEEE Symp. Information Visualization 95*, IEEE CS Press, Los Alamitos, Calif., 1995, pp. 51-58.
4. H. Small, "Visualizing Science by Citation Mapping," *J. Am. Soc. Information Science*, vol. 50, no. 9, 1999, pp. 799-813.
5. K. Fairchild, S. Poltrock, and G. Furnas, "SemNet: Three-Dimensional Graphic Representations of Large Knowledge Bases," *Cognitive Science and Its Applica-*

tions for Human-Computer Interaction, R. Guidon, ed., Lawrence Erlbaum Associates, Hillsdale, N.J., 1988, pp. 201-233.
6. G.S. Davidson et al., "Knowledge Mining with VxInsight: Discovery through Interaction," *J. Intelligent Information Systems*, vol. 11, no. 3, 1998, pp. 259-285.
7. R.W. Schvaneveldt, F.T. Durso, and D.W. Dearholt, "Network Structures in Proximity Data," *The Psychology of Learning and Motivation*, G. Bower, ed., Academic Press, San Diego, Calif.,1989, pp. 249-284.
8. C. Chen, "Generalised Similarity Analysis and Pathfinder Network Scaling," *Interacting with Computers*, vol. 10, no. 2, 1998, pp. 107-128.
9. C. Chen, "Visualising Semantic Spaces and Author Co-citation Networks in Digital Libraries," *Information Processing & Management*, vol. 35, no. 2, 1999, pp. 401-420.
10. S. Lawrence, C.L. Giles, and K. Bollacker, "Digital Libraries and Autonomous Citation Indexing," *Computer*, June 1999, pp. 67-71.

Chaomei Chen *is a reader in the Department of Information Systems and Computing and director of the VIVID Research Center at Brunel University. His research interests are information visualization, virtual environments, human-computer interaction, and hypermedia. He received a PhD in computer science from the University of Liverpool. He is a member of the ACM, the IEEE Computer Society, the Information Visualization Society, and the American Society for Information Science. Contact him at chaomei.chen@brunel.ac.uk.*

Ray J. Paul *is a professor of simulation modeling and director of Research for the Centre for Applied Simulation Modeling and the Centre for Living Information Systems Thinking in the Department of Information Systems and Computing at Brunel University. His research interests are simulation modeling processes, software environments for simulation modeling, and information systems development. He received a PhD in operational research from Hull University. He is a member of the ACM, the IEEE Computer Society, the Operational Research Society, the British Computer Society, and the Society for Computer Simulation. He is the European Chapter chairman of ACM SIGSIM. Contact him at ray.paul@brunel.ac.uk.*