Editors: Jim X. Chen, jchen@cs.gmu.edu
R. Bowen Loftin, bloftin@odu.edu

# A PICTURE FROM A THOUSAND WORDS

By André Skupin

ANYONE ENGAGED IN SCIENTIFIC WORK KNOWS HOW DIFFICULT IT IS TO KEEP UP WITH THE LATEST SCIENTIFIC ADVANCES. THE OVERWHELMING VOLUME OF LITERATURE—RELATED TO THE GROWING NUMBER OF PEER-REVIEWED PRINT outlets and Internet-based sources—combined with cutting-edge science's increasingly interdisciplinary character means that scientists, educators, students, policy makers, and funding agencies need new methods to understand the structure and development of knowledge domains. To address this information glut, I developed a visualization methodology that combines approaches from information science, computer science, and geography. In this article, I present results of a visualization of the geographic knowledge domain based on several thousand conference abstracts as well as a visualization of search results from a research grants database.

## From Text Retrieval to Text Visualization

Disseminating scientific ideas and research results primarily relies on text written in natural language. Traditionally, recognizing structures and relationships in and among different text sources has required time-consuming manual approaches. Computational text analysis speeds up these processes by allowing, for example, keyword search and document-to-document content comparisons. To do this, we must transform text documents into a structured form amenable to computation.

Underlying many text search engines is the vector space model,[1] in which each document is represented as a vector with dimensions corresponding to keywords or terms and numerical weights expressing each term's importance in the document. Another popular technique—especially in the analysis of peer-reviewed scientific literature—is to construct a citation network in which documents become nodes and citations become directed links. To be applicable to a broad range of documents, my visualization methodology constructs a vector space model to analyze implicit document relationships based on shared content, instead of the citations provided by explicit links.

My system preprocesses input data (including removing unwanted stop words, such as articles and prepositions) using standard information-science methods (see the top row in Figure 1). It then constructs a vector space model from the remaining vocabulary and uses document vectors to train a self-organizing map (SOM).[2] This is a special type of artificial neural network that orders neurons in a certain manner, typically to form a 2D lattice. When dealing with a small number of neurons, we can view the SOM as a clustering tool—for example, a 3 × 3 SOM would lead to a nine-cluster solution. However, if we train a SOM with many neurons, we can map out topological relationships among large numbers of $n$-dimensional vectors in the 2D display space. When dealing with document vectors, we can represent individual documents with 2D point geometry. We can visualize the trained SOM in numerous ways, including the delineation of $n$-dimensional cluster structures.

A SOM's 2D form suggests a possible relevance of traditional geographic principles and cartographic techniques, which are mostly geared toward 2D planar representations of geographic reality. My methodology uses commercial geographic information systems (GIS) software (ArcGIS by Environmental Systems Research Institute; www.esri.com) to represent the 2D neuron lattice and various geometric configurations derived from it. The eventual visualization requires many transformations, including clustering of $n$-dimensional neuron vectors and determining appropriate label terms for those clusters. These transformations are linked via loose coupling—that is, file exchange between different components—because there are no existing integrated solutions.[3,4]

## Visualizing Conference Abstracts

Many scientific disciplines hold regular meetings that represent the full range of activities related to the over-

Copublished by the IEEE CS and the AIP 1521-9615/04/$20.00 © 2004 IEEE

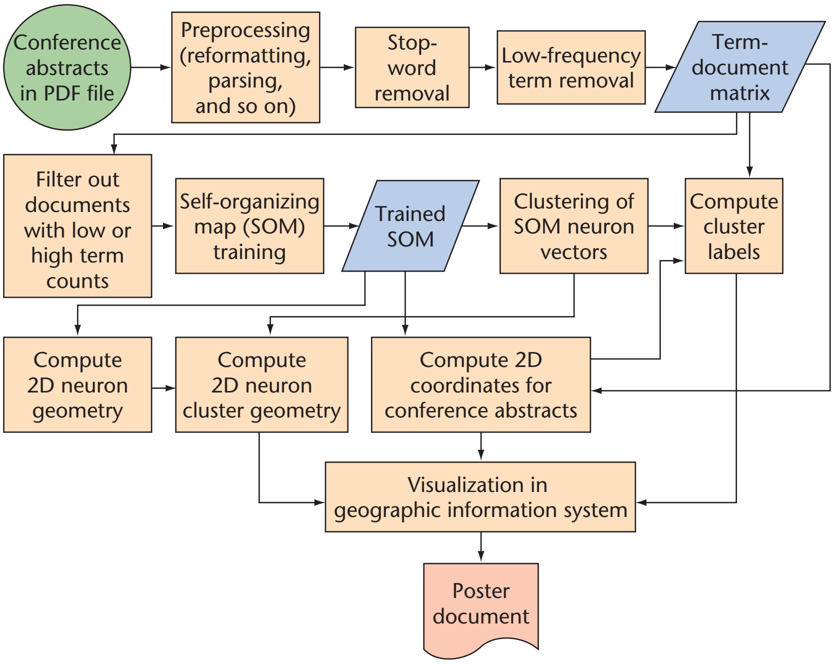
Conference abstracts in PDF file
Preprocessing (reformatting, parsing, and so on)
Stop-word removal
Low-frequency term removal
Term-document matrix
Filter out documents with low or high term counts
Self-organizing map (SOM) training
Trained SOM
Clustering of SOM neuron vectors
Compute cluster labels
Compute 2D neuron geometry
Compute 2D neuron cluster geometry
Compute 2D coordinates for conference abstracts
Visualization in geographic information system
Poster document

**Figure 1. Methodology for deriving a spatialization from a set of conference abstracts. I employed approaches from information science and computer science to preprocess text data and train a neural network. Various geometric transformations then lead up to eventual visualization in a geographic information system. (Courtesy of the National Academy of Sciences.[3])**

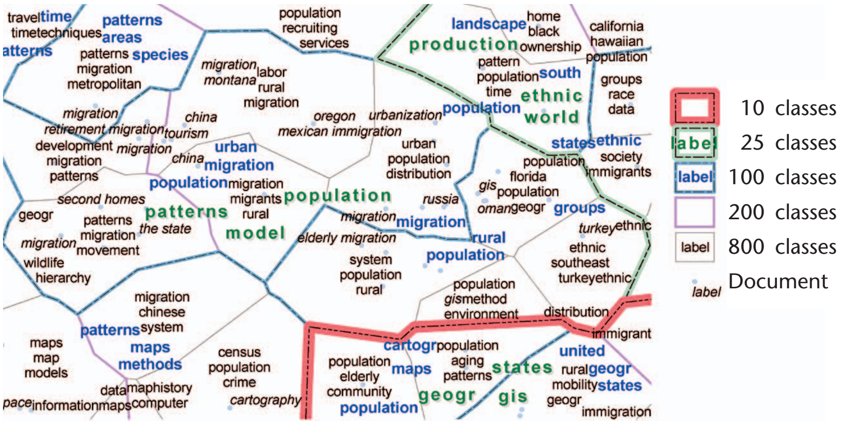

**Figure 2. Portion of a spatialization of conference abstracts. Five levels of a hierarchical clustering solution are delineated simultaneously. (Copyright 2004 National Academy of Sciences USA.[3])**

arching knowledge domain, instead of being devoted to individual, highly focused topics. Within the geography field, the Association of American Geographers' (AAG; www.aag.org) annual meeting fulfills this function. With several thousand presentations given at each year's meeting, it's a prime event for gauging the discipline's current state and potential direction.

The large number of presentations necessitates that the printed conference program lists only presentation titles; a CD-ROM contains the presentation abstracts (approximately 250 words each). I believed that this meeting was a likely candidate for attempting a computational analysis of the conference presentations and thus of the entire geographic discipline. Furthermore, if the computational result were to be in a visualization form, the knowledge domain might become accessible to many users.

We've used geographic maps in research, education, administration, and other areas for centuries, and many maps have aesthetic values. We might use knowledge-domain visualizations in similarly diverse circumstances. For example,

- in a college-level introductory course, students might see a discipline's overall structure;
- researchers engaged in interdisciplinary scientific work might gain quick understanding of an area outside their own core domain; and
- funding agencies could use visualizations to spot emerging research trends.

To demonstrate this, I processed 2,220 abstracts from the AAG's 1999 meeting. Cartographic considerations, especially in symbology choices and scale issues, influenced the eventual visualization's design (see Figures 2 and 3). I designed Figure 2 to evoke the notion of an administrative subdivision consisting of countries, provinces, counties, and so on. Hierarchical clustering of neurons places them into a nested cluster structure. To the map's users, the structure manifests itself through a visual hierarchy, with varying line thickness of cluster boundaries, different font sizes for cluster labels, and matching colors for lines and labels at the same cluster level.

The map-space tessellation that hierarchical clustering provides has obvi-

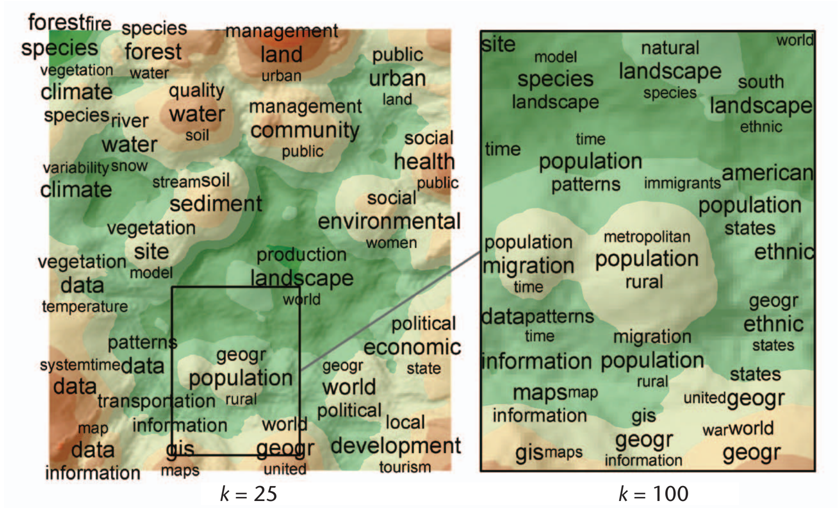

**Figure 3. Spatialization by means of a term-dominance landscape, combined with scale-dependent labeling to support semantic zooming. Labeling is based on (a) *k* = 25 and (b) *k* = 100 *k*-means cluster solutions. (Copyright 2004 National Academy of Sciences USA.[3])**

**Figure 4. Geographic conference abstract poster. Geography faculty members discuss their own knowledge domain while facing a large-format visualization derived from several thousand conference abstracts.**

ous advantages. Map users readily recognize the main intended metaphor (an administrative regions map). The nested structure provides simultaneous display of different cluster levels and avoids overly complex line work. In an interactive setting, it easily can implement scale changes in a cognitively plausible form. For example, when changing from a 25- to a 100-cluster solution during zoom-in, the higher-level cluster boundaries manifest themselves again as lower-level boundaries and thus provide landmark-type features during multiscale exploration.

However, hierarchical clustering's consistency and strictness means that feature space subdivision is far from optimal. Of course, the purpose of clustering is not to find the single best partition; instead, we are looking for a partition that is computationally adequate and cognitively supportive of multiscale exploration. If we use other, more optimal, computational methods, then major adjustments to visualization methods could be necessary. For example, *k*-means clustering tends to produce feature space partitions that more closely follow existing high-dimensional structures, compared with hierarchical clustering.[3] However, different *k* solutions (for example, *k* = 25 versus *k* = 100) compute independently, which means that simultaneous display of multilevel boundaries would be too complex visually. To address this, I experimented with providing feature space geometry as an elevation field derived through a term-dominance landscape. I overlay *k*-means clusters with boundary lines switched off, and cluster locations approximated by placing cluster labels near respective centroids (see Figure 3). Thus, multiscale exploration is based on recognition of major mountain ranges, peaks within these ranges, and so forth.

Generally, researchers have considered interactivity to be crucial to any information visualization.[5] One of my goals was to demonstrate that such a view of information visualization could be too narrow. As an example, I use geographic maps in various forms of geographic discourse. First, when soliciting citizen input about land-use zoning changes during a town-hall meeting, the geographic depiction of existing and planned land use doesn't serve only as illustration. These maps also enable (by establishing a shared geographic reference, such as a neighborhood's extent) and then shape discourse among decision makers and those potentially affected. Also, consider how a large wall map helps frame the introduction to a certain geographic problem in a college-level geography course. While such visualizations are static, they still encourage interaction among viewers and with the map itself, which might not physically change, but nevertheless allows fresh discovery and reflection of structures and relationships every time someone views it.

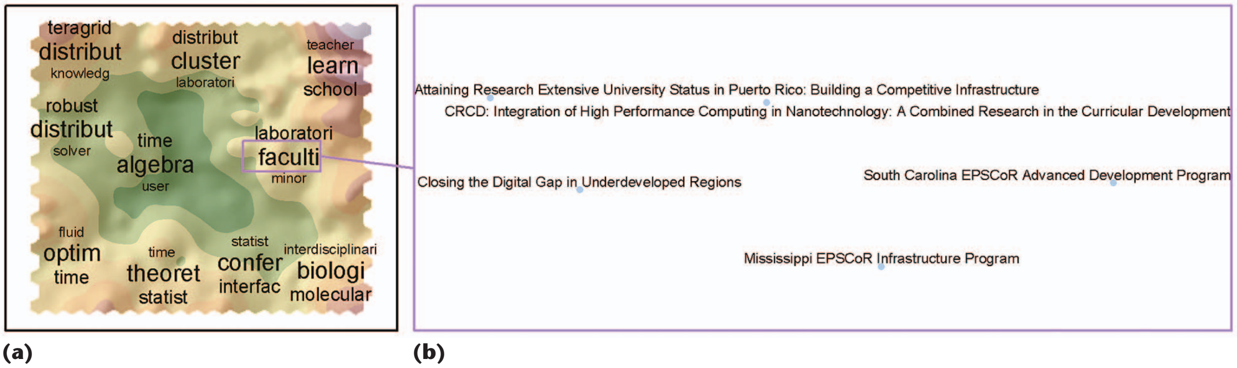
teragrid
distribut
knowledg
distribut
cluster
laboratori
teacher
learn
school
robust
distribut
solver
time
algebra
user
laboratori
faculti
minor
fluid
optim
time
time
theoret
statist
statist
confer
interfac
interdisciplinari
biologi
molecular
(a)
Attaining Research Extensive University Status in Puerto Rico: Building a Competitive Infrastructure
CRCD: Integration of High Performance Computing in Nanotechnology: A Combined Research in the Curricular Development
Closing the Digital Gap in Underdeveloped Regions
South Carolina EPSCoR Advanced Development Program
Mississippi EPSCoR Infrastructure Program
(b)

**Figure 5. Visualization of a search for documents containing "computing" and "sciences" in the US National Science Foundation's online database. (a) The global map constructed from all matching grants and (b) specific locations and titles of grants within a smaller region.**

To demonstrate that knowledge-domain visualizations could serve similar purposes and be used in similar ways, I experimented with poster-size prints derived from the same data set of geographic conference abstracts. Figure 4 shows one possible scenario. A group of geography professors are discussing (possibly rediscovering) their own discipline, while faced with a visualization that simultaneously shows broad subdivisions and finer structures within the geographic knowledge domain.

## Visualizing Grant Abstracts

Organizations engaged in providing funds for scientific research receive thousands of grant proposals each year. To those applying, it is often difficult to clearly understand the kinds of topics an agency supports and how topical support structures might have changed over time. The growing trend toward interdisciplinary research makes it more important to gain a broad perspective. Likewise, funding agencies want to track the relative success of previously funded work, to assess relative merit of new proposals and detect patterns in the impact of different researchers and research approaches. In the corporate arena, venture capitalists might want to put a startup company's vision into the context of overarching trends as reflected by recently funded research. In all of these scenarios, research-grant information visualization could be an important tool for decision making. Indeed, this has become a growing area of investigation among information scientists and digital library specialists.[6]

To develop a grant abstracts visualization example, I started by querying the US National Science Foundation site (www.nsf.gov) for grants containing certain terms (in this case, "computing" and "sciences") and parsed the query response into an XML file. This was input to a process very similar to the one Figure 1 shows. One important difference was that I used stemming to reduce all words to root forms.

After filtering out low-frequency terms (to reduce overall dimensionality in the vector space model) and also removing abstracts with very low term counts (that is, very short abstracts with too little substantive content), 162 grant abstracts went into the visualization. Following SOM training, I computed *k*-means cluster solutions and a term-dominance landscape and used ArcGIS for the final symbolization (Figure 5).

Despite the small number of input documents, meaningful structures emerge. In the upper left and center of Figure 5a are core computing topics (distributed and cluster computing); in the lower left are more applied, computing topics, such as fluid and temporal modeling of various kinds and optimization procedures. Applications of computing to biological topics occupy a distinct location in the bottom right-hand corner. Finally, the upper-right quadrant is occupied by educational topics.

Figure 5b's zoomed-in view shows the region labeled "faculti, laboratori, minor." The curious word endings hark back to the stemming algorithm—for example, "minor" derived from such terms as "minority" or "minorities". Grants that address historically evolved inequities in scientific computing mostly occupy this region. Titles such as "Closing the Digital Gap in Underdeveloped Regions" or "Attaining Research Extensive University Status in Puerto Rico: Building a Competitive Infrastructure" are indicative of this. The Experimental Program to Stimulate Competitive Research (EPSCoR; www.ehr.nsf.gov/epscor) program aims to increase research resources for states with historically small federal funding, represented here by proposals from Mississippi and South Carolina.

Visualization of scientific writing for science education, management, and funding is an interdisciplinary endeavor that I expect will gain importance in the future. In May 2003, the National Academy of Sciences hosted a Sackler Symposium dedicated solely to the mapping of knowledge domains.[7] Participants represented the full spectrum of activities in this area, from those working on fundamental techniques for information extraction and organization to large-scale implementations of such techniques for domain-specific repositories. An example of the latter is the application of a machine-learning approach to an online repository of several hundred thousand articles dealing mostly with physics (www.arxiv.org).[8] Several visual techniques also were discussed, ranging from the type of visualizations shown here to highly interactive methods.

Today, visualization has emerged as one of the principal strategies for dealing with data glut. On the other hand, the continued dominance of ranked lists as the primary output from search engines—as opposed to holistic, map-like displays—demonstrates that this approach still has a long way to go. I've shown that one secret for compelling visualizations might be the deliberate combination of intense computation coupled with traditional cartographic design approaches. In this context, many research issues might benefit from ongoing cartographic and geographic involvement. Fundamental questions remain to be answered about the use of map metaphors, especially regarding their cognitive plausibility when dealing with non-geographic, high-dimensional data. We also need to learn how to balance the number of items to be displayed with the available display space, drawing inspiration from methods of abstraction used in geographic maps. As we implement semantic zooming, the determination and display of appropriate label terms remains another important issue. More than anything else, the prime factor contributing to the ongoing development and success of knowledge-domain visualizations will be the continued nurture of cross-fertilization and collaboration among various scientific disciplines. CiSE

## Acknowledgments

The research presented here is partially supported by the Louisiana Board of Regents Support Fund, grant #LEQSF (2002-05)-RD-A-34. Aga Shirazi, University at Buffalo, assisted with parsing and preprocessing the US National Science Foundation grant abstracts.

## References

1. G. Salton, *Automatic Information Organization and Retrieval*, McGraw-Hill, 1968.
2. T. Kohonen, *Self-Organizing Maps*, Springer-Verlag, 1995.
3. A. Skupin, "The World of Geography: Visualizing a Knowledge Domain with Cartographic Means," *Proc. Nat'l Academy Sciences*, vol. 101, Suppl. 1, 2004, pp. 5274–5278.
4. A. Skupin, "A Cartographic Approach to Visualizing Conference Abstracts," *IEEE Computer Graphics and Applications*, vol. 22, no. 1, 2002, pp. 50–58.
5. S.K. Card, J.D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, 1999.
6. K.W. Boyack and K. Börner, "Indicator-Assisted Evaluation and Funding of Research: Visualizing the Influence of Grants on the Number and Citation Counts of Research Papers," *J. Am. Soc. Information Science Technology*, vol. 54, no. 5, 2003, pp. 447–461.
7. R.M. Shiffrin and K. Börner, "Mapping Knowledge Domains," *Proc. Nat'l Academy Sciences*, vol. 101, Suppl. 1, 2004, pp. 5183–5185.
8. P. Ginsparg et al., "Mapping Subsets of Scholarly Information," *Proc. Nat'l Academy Sciences*, vol. 101, Suppl. 1, 2004, pp. 5236–5240.

**André Skupin** is an associate professor in the Department of Geography at the University of New Orleans. His research interests include text document visualization, geographic visualization, cartographic animation and hypermedia, and cartographic generalization. He has a Dipl.-Ing. in cartography from the Technical University, Dresden, Germany, and a PhD in geography from the State University of New York at Buffalo. He did graduate research with the National Center for Geographic Information and Analysis (NCGIA) and has worked in the GIS industry in the US, Germany, and South Africa. Contact him at askupin@uno.edu.

**Submissions:** Send one PDF copy of articles and/or proposals to Francis Sullivan, Editor in Chief, fran@super.org. Submissions should not exceed 6,000 words and 15 references. All submissions are subject to editing for clarity, style, and space.

**Editorial:** Unless otherwise stated, bylined articles and departments, as well as product and service descriptions, reflect the author's or firm's opinion. Inclusion in *CiSE* does not necessarily constitute endorsement by the IEEE, the AIP, or the IEEE Computer Society.

**Circulation:** *Computing in Science & Engineering* (ISSN 1521-9615) is published bimonthly by the AIP and the IEEE Computer Society. IEEE Headquarters, Three Park Ave., 17th Floor, New York, NY 10016-5997; IEEE Computer Society Publications Office, 10662 Los Vaqueros Circle, PO Box 3014, Los Alamitos, CA 90720-1314, phone +1 714 821 8380; IEEE Computer Society Headquarters, 1730 Massachusetts Ave. NW, Washington, DC 20036-1903; AIP Circulation and Fulfillment Department, 1NO1, 2 Huntington Quadrangle, Melville, NY 11747-4502. Annual subscription rates for 2004: $42 for Computer Society members (print only) and $48 for AIP society members (print plus online). For more information on other subscription prices, see www.computer.org/subscribe or https://www.aip.org/forms/journal_catalog/order_form_fs.html. Computer Society back issues cost $20 for members, $96 for nonmembers; AIP back issues cost $22 for members.

**Postmaster:** Send undelivered copies and address changes to Circulation Dept., *Computing in Science & Engineering*, PO Box 3014, Los Alamitos, CA 90720-1314. Periodicals postage paid at New York, NY, and at additional mailing offices. Canadian GST #125634188. Canada Post Publications Mail Agreement Number 40013885. Return undeliverable Canadian addresses to 4960-2 Walker Rd., Windsor, ON N9A 6J3. Printed in the USA.

**Copyright & reprint permission:** Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limits of US copyright law for private use of patrons those articles that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Dr., Danvers, MA 01923. For other copying, reprint, or republication permission, write to Copyright and Permissions Dept., IEEE Publications Administration, 445 Hoes Ln., PO Box 1331, Piscataway, NJ 08855-1331. Copyright © 2004 by the Institute of Electrical and Electronics Engineers Inc. All rights reserved.