Mapping knowledge domains

Richard M. Shiffrin*† and Katy Börner‡

*Psychology Department and ‡School of Library and Information Science, Indiana University, Bloomington, IN 47405

The term "mapping knowledge domains" was chosen to describe a newly evolving interdisciplinary area of science aimed at the process of charting, mining, analyzing, sorting, enabling navigation of, and displaying knowledge. This field is aimed at easing information access, making evident the structure of knowledge, and allowing seekers of knowledge to succeed in their endeavors. Although thousands of years old, this area has undergone a sea change in the last 15 years, a change fostered by an explosion of the amount of information available, the accessibility of that information due to electronic storage, and the new techniques of analysis, retrieval, and visualization that are made possible by vast increases in computational storage capacity and processing speed and power. Many of us are so involved in the new ways of accessing knowledge that we have forgotten how recent is the change to computerized knowledge retrieval with search engines operating on the World Wide Web. Remarkable as these changes are to date, they are only a hint of the transformation to come. The Arthur M. Sackler Colloquium on Mapping Knowledge Domains, held at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA, May 9–11, 2003, was designed to showcase the ongoing developments in this transformation and provide pointers toward the directions it will move.

The changes that are taking place profoundly affect the way we access and use information. Scientists, academics, and librarians have historically worked hard to codify, classify, and organize knowledge, thereby making it useful and accessible. The day is fast approaching when all this knowledge will be coded electronically, but mixed in a vast and largely disorganized and often unreliable sea of mostly recent information. Fishing this sea for desired information is presently no easy task and will continue to increase in difficulty. However, the speed and power of modern computation gives hope that this daunting task can be accomplished. In addition, and perhaps even more important, the new analysis techniques that are being developed to process extremely large databases give promise of revealing implicit knowledge that is presently known only to domain experts, and then only partially.

Some of these techniques are now being applied in science, aiming to identify and organize research areas according to experts, institutions, grants, publications, journals, citations, text, and figures; discover interconnections among these; establish the import of research; reveal the export of research among fields; examine dynamic changes such as speed of growth and diversification; highlight economic factors in information production and dissemination; find and map scientific and social networks; and identify the impact of strategic and applied research funding by government and other agencies. The new techniques support and complement human judgment. They dramatically speed up achievements formerly reached solely by human effort and provide new results that could not have been reached by humans unaided. As the flood of new and disorganized information continues to crest, the new tools are increasingly critical for the growth of scientific research, and indeed for the functioning of modern society.

The importance and fundamental nature of these new ways of interacting with information, and accessing knowledge, have led to considerable interest in for-profit applications. As a result, many of the algorithms and software developed in this field are proprietary. Users are given the end products, such as a list of potentially useful websites or a visual map, without much knowledge concerning the conceptual basis and technical implementation of the underlying algorithms. The desire to promote a deeper understanding therefore led us to include leading researchers not only from academia and government, but also from businesses such as Google and Microsoft.

We thought it would prove useful and interesting if some of the techniques used to map knowledge were applied to the contents of PNAS itself. Thus, we arranged for registered participants to have access to an electronic compilation of the full text documents from PNAS covering January 7, 1997, to September 17, 2002 (148 issues containing some 93,000 journal pages). The time between the first availability of this data set and the deadline for submissions was rather short; nonetheless, several of the contributors analyzed this set, with results that provide interesting directions for future research.

The value of mapping knowledge domains of course extends well beyond the bounds of information science or the PNAS journal, to scientists, researchers, governmental institutions, industry, and members of society generally. It should be emphasized that, although the extraction and organization of knowledge may form the scientific core of this field, the results will be of little use unless the user can understand and interact with the mapping systems. Knowledge typically is organized along many thousands of dimensions, but a map with thousands of dimensions cannot be used effectively by humans. For this reason, domain visualizations and the ability to interact with knowledge and view it from a variety of perspectives play a critical role. The results of algorithms used to extract and organize relevant data can be displayed in many complementary ways. For example, maps might depict major researchers, most cited articles and books, articles too new to receive many citations but with contents that point to emerging trends, articles organized into topic trees (by content, citations, and authors), and grants awarded by topic. Other maps might depict changes over time. Such techniques hold out the promise that the user will be able not only to visualize a few nearby trees in the forest of knowledge, but also to understand the entire landscape. If these techniques can be made to operate effectively, they may well change the way that science is conducted and the way the business of the world is carried out.

Achieving such results requires tools from diverse areas of science: ways to analyze truly enormous amounts of data and extract meaningful results; ways to sort and cluster information

This paper serves as an introduction to the following papers, which result from the Arthur M. Sackler Colloquium of the National Academy of Sciences, ''Mapping Knowledge Domains,'' held May 9–11, 2003, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA.

†To whom correspondence should be addressed. E-mail: shiffrin@indiana.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0307852100

by similarity and importance; ways to identify close and distant interconnections that are not immediately obvious (especially when terminology differs); ways to display large amounts of information that lie along multiple dimensions so that the user can properly interpret the results and guide further exploration; ways to design interactive interfaces; and ways to analyze the structure of the database itself. Examples of many of these are represented in the articles of this special issue, and additional examples were presented at the colloquium (slides and audio files of the talks and a video of the poster presentations are accessible at http://vw.indiana.edu/sackler03/). In the long run, the promise of this field will not be realized without dynamically interactive systems; several of our participants have been developing such systems, but the pages of a journal do not, unfortunately, afford access to dynamic systems.

Overview of Contributions

This special PNAS issue contains three articles that set the stage by providing general coverage of methods, techniques, and practices: an analysis of knowledge extraction from the World Wide Web by Monika Henzinger and Steve Lawrence (Google research); an analysis, correlation, and mapping of paper and grant data to assess research by Kevin W. Boyack (Sandia National Laboratories, Albuquerque, NM); and an analysis of scientific collaboration networks by Mark Newman (University of Michigan, Ann Arbor).

Several articles address methods to extract and organize information from large unstructured databases: Simon Dennis presents an unsupervised method to extract propositional information from a "tennis article" database and answer questions about information implicit in the data. Three articles present methods to organize databases in terms of the semantic similarity of the contents and to apply the methods to the PNAS database. Tom Landauer, Darrell Laham, and Marcia Derr use "Latent Semantic Analysis." Elena Erosheva, Stephen Fienberg, and John Lafferty use a form of mixed-membership model, and Tom Griffiths and Mark Steyvers present another form they term the "Topics Model" (both are a generalization of "Latent Dirichlet Allocation"). Paul Ginsparg, Paul Houle, Thorsten Joachims, and Jae-Hoon Sul classify research areas inherent in a large text database of physics articles by using a "support vector machine."

Other articles assume a knowledge database represented or representable as a graph structure. Dennis Wilkinson and Bernardo A. Huberman present a stochastic network partitioning procedure that extends the "Girvan–Neuman" method to large, complex graphs to account for nodes that belong to several clusters. They use it to identify communities of genes related to colon cancer. The method could as well be used to determine communities of papers or authors from paper-citation or coauthor networks.

Most data sets evolve over time, and several articles address ways to track dynamic changes in structure. One approach analyzes the changes in users' interactions with the database. The article by John Hopcroft, Omar Khan, Brian Kulis, and Bart Selman applied a new clustering algorithm to the NEC CiteSeer database that would identify real changes in structure without identifying changes due to random and small perturbations. Jonathan Aizen, Daniel Huttenlocher, Jon Kleinberg, and Antal Novak present a stochastic method to analyze the dynamics of item popularity (web traffic) on the Internet Archive and use it to identify points of significant change in real world events.

Given the complexity and nonlinearity of the structure and evolution of, for example, article networks, coauthor networks, and web page graphs, those who wish to mine such networks for knowledge must understand the processes by which such networks evolve. Filippo Menczer introduces a mixture model that grows web page or paper-citation networks on the basis both of based existing interlinkages (popularity) and of the content of the papers and web pages, thereby reproducing network "degree" and content similarity distributions. Katy Börner, Jeegar Maru, and Robert Goldstone present a simple process model that simultaneously grows coauthor and paper-citation networks; the statistical and dynamic properties of the simulated network data are validated against a 20-year PNAS data set.

Methods to display and visually explore the results of large-scale database analyses (i.e., visualization techniques) are of critical importance, and these can benefit from centuries of work in geographic metaphors and cartographic techniques. André Skupin reviews major cartographic principles and by way of example produces a large-format map-like knowledge-domain visualization. Alan MacEachren, Mark Gahegan, and William Pike combine geographic visualization techniques and concept mapping to design a tool that helps individual researchers describe the process of knowledge construction and enables teams of collaborators to synthesize common concepts. Ketan Mane and Katy Börner identify and correlate highly frequent and highly bursty words in the PNAS database to visualize the structure and evolution of major research topics over time. Steven Morris and Gary Yen introduce Crossmaps, a technique that can be applied to visualize multiple, overlapping relations among documents such as author collaboration groups vs. topics or research fronts covered by those documents. Howard White, Xia Lin, Jan Buzydlowski, and Chaomei Chen present a tool (and apply it to the PNAS database) that automatically and rapidly generates small-scale, "local" pathfinder networks and self-organizing maps as interfaces for document retrieval. The interfaces show co-cited authors or co-occurring subject headings that can be explored interactively. Chaomei Chen presents a technique to interrelate and visually present network structures of, say, paper-citation or coauthor networks, generated for different time slices.

Opportunities and Challenges

The increasing flood of digitally available data demands the development of sophisticated tools of analysis and display, but the tools are limited by the quality of the data. All of the research described in this special issue requires high-quality data that are both accessible and available in a usable and common format. A good deal of "preprocessing" was in most cases required to transform data to reach this state. We hope to see in the near future the development of tools to take information in different and noisy formats and convert it to a common format, or analyze noisy and inconsistent data directly.

Because present analysis requires clean data, it is often necessary to make use of proprietary databases. There is often a cost of access, a fact that contributes to an increasing "information divide." There are of course efforts to move beyond private information (such as Medline, the ACM digital library, or the Physics E-print Archive), but these are currently unconnected.

Such factors are slowing the development of truly global and freely accessible maps of science, or of general knowledge, but we hope and believe that day will arrive. The Sackler Colloquium on Mapping Knowledge Domains and the present articles that derived from that colloquium provide provocative glimpses of that future day.

Many of our colleagues participated in the organization of this colloquium. We thank the members of our organizing committee: Kevin Boyack, Sandia National Laboratories; Chaomei Chen, Drexel University; Susan Dumais, Microsoft Corporation; Jon Kleinberg, Cornell University; Thomas K. Landauer, University of Colorado; and Josh Tenenbaum, Massachusetts Institute of Technology. We greatly appreciate the time and effort of our additional reviewers; these included many of the authors of the present articles, and also Rex G. Cammack, Blaise Cronin, Tom Erickson, Eugene Garfield, Beth Hetzler, Peter Ingwersen, Grant Lewison, Francis Narin, Sidney Redner, Ben Shneiderman, and Colin Ware. We also thank Kevin Boyack, Jason Baumgartner, and Ketan Mane for processing and managing access to the Colloquium's PNAS database. Special thanks must also go to Eugene Garfield, the "father" of this field, who gave the keynote at the Colloquium and participated actively in the colloquium, adding greatly to its value. The 5-year PNAS full text data set was provided by Diane M. Sullenberger, executive editor, PNAS. The 20-year PNAS citation data set was extracted from *Science Citation Index Expanded* [Institute for Scientific Information, Inc. (ISI), Philadelphia, PA; Copyright ISI]. All rights reserved. No portion of these data may be reproduced or transmitted in any form or by any means without the prior written permission of ISI.