Mapping Scientific Frontiers

Chaomei Chen

Mapping Scientific Frontiers

The Quest for Knowledge Visualization

Second Edition

Springer

Chaomei Chen
College of Information Science and Technology
Drexel University
Philadelphia, Pennsylvania
USA

ISBN 978-1-4471-5127-2 ISBN 978-1-4471-5128-9 (eBook)
DOI 10.1007/978-1-4471-5128-9
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2013944066

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

Mapping science at first glance appears to be an oxymoron: how can we map something as abstract as science? Scientific knowledge seems to occupy an intellectual realm which can only be glimpsed by the mind of the highly trained specialist. Yet this book demonstrates that the discipline of science mapping has been going on for a number of years and is indeed beginning to flourish with new results and insights. In this endeavor we may only be in the stage of the early explorers who drew out the first crude geographic maps of the then known physical world. We might argue that science mapping is simply the logical progression of map making from the physical to the intellectual world.

It is not necessary to make the case for the importance of science in the modern world and the object of our map making efforts, even though our society has at times minimized the role of scientific research, neglected to provide adequate funding and support, and attempted to retard its educational programs. What more important artifact of human intelligence could we focus on than the current scientific landscape? Of course science can be used for good or ill, and science confers incredible power on those who master its principles. On occasion individual scientists will abuse the trust we place in them in their quest for recognition, as Chaomei Chen documents here in his study of "retractions" of scientific articles. But despite these aberrations, science is the gateway to understanding of our place in the universe, and the foundation of our social and economic well-being.

Despite the fact that the language we use to describe science is replete with spatial metaphors such as "field" and "area" of research, when we actually go about trying to create a map of science, we soon realize that the procedures used to make geographic maps no longer apply. We must deal with the abstract relations and associations among entities such as scientific ideas, specialties, fields or disciplines whose very existence may be open to debate. Thomas Kuhn described these shifting historical disciplines in a revealing interview: "Look ... you must not use later titles for [earlier] fields. And it's not only the ideas that change, it's the structure of the disciplines working on them." (Kuhn 2000, p. 290) Here we realize that science in any historical period, and indeed the current period, is a *terra incognita*. Are we justified in seeking a spatial representation of such abstract and perhaps hypothetical entities? Are our brains hardwired to take what is relational and project it in real space?

Perhaps science mapping is difficult to grasp because three conceptual steps are required before it makes sense, two of which involve some degree of mathematical manipulation. First a unit of analysis must be chosen which should comprise the elementary particles of our science universe. Secondly a measure of association between the units must be defined. Thirdly, a means must be found for depicting the units and their relations in a low dimensional space (usually two dimensions) that can be perceived by humans. Once these intellectual leaps are made science mapping seems natural and even inevitable.

The established scholarly disciplines of the history of science, sociology of science and philosophy of science have long been regarded as providing the essential tools for understanding the origin and evolution of scientific thought, its social institutions, and its philosophical underpinnings. With the exception of some early work in the sociology of science, the methods used by these disciplines have been largely qualitative in nature. History of science has dealt largely with constructing narratives using the methods of general history, and philosophy of science with logical foundation and epistemology. Sociology of science, as exemplified by one of its founders Robert Merton, was both strongly theoretical and open to the use of quantitative evidence. This approach was also taken up by early social network researchers who studied so-called invisible colleges. In the 1970s, however, sociology of science turned away from quantitative methods to embrace more radical social theories of science, such as the social construction of scientific knowledge, and largely abandoned the quantitative approach of the earlier sociologists.

More recently, primarily as the result of the availability of large databases of scientific publications and citations and partly a reaction against constructivist sociology, a discipline emerged which has opened up a new way to study the evolution of science. This field has been called variously scientometrics, informetrics and bibliometrics. These terms reflect not only the focus on quantitative methods upon which it is built, but also its origins in what was once called library science. It cannot be claimed that the upstart discipline has achieved acceptance in the academic world, particularly on the part of the more established disciplines, and institutionalization in the form of university programs and academic positions has only just begun. Critics of scientometrics have claimed that a focus on the scientific literature as its primary source material too severely limits the data on which studies of science can be based. On the other hand, the increasing availability of the full text of scientific papers in computer readable formats, opens up many new types of data for analyses which when used in tandem with the standard on-line databases goes far beyond what has been possible using the standard indexes alone. Combined with software packages such as Chaomei Chen has pioneered, a powerful new tool for the analysis of science has come into being. There is every indication that the new field is here to stay and is exerting more and more influence on policy, even though a rapprochement and integration with the traditional fields of history, sociology and philosophy is probably a long way off.

Chaomei Chen's book is important because it builds on many of the concepts and findings of history, sociology and philosophy of science, but at the same time adds a new dimension. As an example of the power of the new methods the skeptical reader should consult chapter eight presenting a case study on recent work on induced pluripotent stem-cells which shows how mapping can inform historical studies as well as assist medical researchers to get an overview of a research area. Here we see the strength of the new methods for exploring and tracking the internal structure of revolutionary developments in contemporary science.

His book also draws on an even broader disciplinary framework from computer to information science and particularly information visualization. In the first edition Chaomei Chen commented on the disciplines that contribute ideas to science mapping. "Different approaches to mapping scientific frontiers over recent years are like streams running from several different sources .... A lot of work needs to be done to cultivate knowledge visualization as a unifying subject matter that can join several disciplines." (Chen 2003, p. vii) This remains true even today when scientometrics, computer science, and network science continue to evolve in a strangely independent manner yet often dealing with the same underlying data and issues. This may be an inevitable side effect of the barriers between disciplines, but hopefully this book will help bridge these various streams.

As an example of the relevance of history of science, Chaomei Chen comments that the work of Thomas Kuhn was an important backdrop to mapping because one could think of the unfolding of a revolution in science as a series of cross-sectional maps that at some point undergoes a radical structural transformation. Cross sectional thinking is also very much encouraged in the history of science because historians are exhorted to understand the ideas of a historical period by entering its mind-set, "to think as they did" (Kuhn 1977, p. 110), and not interpret older science in terms of our "current" understanding. This is a difficult requirement because once we know that a new discovery or finding has occurred it is extremely difficult for us not to be influenced by it, and our first impulse is to find precursors and antecedents. As analysts we need to take care not to allow the present to distort the past.

As an example of how various cross-currents converge in science mapping we could point out the tension between psychological factors, as exemplified by Kuhn's gestalt switching as a way of looking at conceptual change, and social forces such as collegial networks and invisible colleges. Do social relations determine cognitive relations, or vice versa? In Stanley Milgram's early work (1967) on social networks, human subjects were required to think about what acquaintances their acquaintances had several steps removed. In Don Swanson's work (1987) on undiscovered public knowledge, discoveries are made by seeking concepts that are indirectly related through other concepts that are currently unconnected. Thus the same type of thinking is involved in both the social and intellectual tasks. If we are dealing with words or references as our mapping units, then psychology clearly enters the picture because an author's memory and recall are involved in the associative process. But that memory and recall are also influenced by what authors have seen other authors or colleagues say. If we map individual scientists in their co-author relations, then social factors must come into play but psychological factors also contribute to the selection of co-authors. Thus social and psychological factors are inexorably intertwined in both the social and intellectual structure of science.

The competition in science mapping between the various choices for unit of analysis such as words, references, authors, journals, etc. and the means of associating them such as co-word, co-citation, co-authorship, direct citation, etc. seems to boil down to the types of structures and level of relations we want to observe. To better understand the role of discovery in specialty development we might turn to co-citations because many discoveries are associated with specific papers and authors. On the other hand, if we want to include broader societal, non-scholarly factors then we might turn to co-words which can more readily capture public or political sentiments external to science. Journals, a yet broader unit of analysis, might best represent whole fields or disciplines. Choice of a unit of analysis also depends on the historical period under investigation. Document co-citation is probably not feasible prior to 1900 due to the absence of standard referencing practice. However, name co-mention within the texts of scientific papers and books is still very feasible for earlier periods. It is instructive to try to imagine how we would carry out a co-mention or other kind of mapping for some earlier era, say for scientific literature in the eighteenth century and whether we would be able to identify the schools of thought and rival paradigms active during the period.

Another important issue is the interpretation of maps. We know that the network of associations that underlies maps is hyper dimensional, and that projection in two dimensions is inevitably an approximation and can place weakly related units close together. This argues for the need to pay close attention to the links themselves which give rise to the two dimensional solution in the first place, which we can think of as the neurons of the World Brain (Garfield 1968) we are trying to visualize. Only by knowing what the links signify can we gain a better understanding of what the maps represent. This will involve looking more deeply at the context in which the linking takes place, and seeking new ways of representing and categorizing those relationships, for example, by function or type such as logical, causal, social, hypothetical, metaphorical, etc. One positive development in this direction, as described in the final chapter, is the advent of systems for "visual analytics" that allow us to more deeply probe the underpinnings of maps with the ultimate goal of supporting decision making.

Part of what is exciting about science mapping is that the landscape is continually changing: every year there is a new crop of papers and the structure changes as new areas emerge and existing areas evolve or die off. Some will find such a picture unsettling and would prefer to see science as a stable and predictable enterprise, but as Merton has argued (2004), serendipity is endemic to science, and thus also to science maps. We do not yet know if discovery is in any way predictable, if there are recognizable antecedents or conditions, or whether discovery or creativity can be engineered to happen at a quicker pace. But because discoveries are readily apparent in maps after they occur, we also have the possibility of studying maps for previous time periods to look for structural antecedents.

This is a wide ranging book on information visualization, with a specific focus on science mapping. Science mapping is still in its infancy and many intellectual challenges remain to be investigated and many of which are outlined in the final chapter. In this new edition Chaomei Chen has provided an essential text, useful both as a primer for new entrants and as a comprehensive overview of recent developments for the seasoned practitioner.

SciTech Strategies, Inc. Henry Small

References

Chen C (2003) Mapping scientific frontiers. Springer, London
Garfield E (1968) "World Brain" or Memex? Mechanical and intellectual requirements for universal bibliographic control. In: Montgomery EB (ed) The foundations of access to knowledge. Syracuse University Press, Syracuse, pp 169–196, from http://garfield.library.upenn.edu/essays/v6p540y1983.pdf
Kuhn TS (1977) The essential tension. University of Chicago Press, Chicago
Kuhn TS (2000) The road since structure. University of Chicago Press, Chicago
Merton RK, Barber E (2004) The travels and adventures of serendipity. Princeton University Press, Princeton
Milgram S (1967) The small world problem. Psychol Today 2:60–7
Swanson DR (1987) Two medical literatures that are logically but not bibliographically connected. J Am Soc Info Sci 38:228–233

Preface for the 2nd Edition

The first edition of *Mapping Scientific Frontiers* (MSF) was published over 10 years ago in 2002. Since then, a lot has changed. Social media has flourished to the extent that we have never seen before. News, debates, hoaxes, and scholarly blogs all fight for attention on Facebook (launched in 2004), YouTube (2005), and Twitter (2006), which are made ubiquitously accessible by popular mobile devices such as iPhone (2007) and iPad (2010).

Over the past 10 years, remarkable scientific breakthroughs have been made, for example, Grigori Perelman's proof of the century-old Poincaré Conjecture in 2002, the Nobel Prize winning research on induced pluripotent stem cells (iPSCs) by Shinya Yamanaka and his colleagues since 2006, and the recent discovery of the Higgs Boson in 2012 at CERN.

The big sciences continue to get bigger. Large-scale data collection efforts for scientific research such as the Sloan Digital Sky Survey (SDSS) (2000–2014) in astronomy represent one of many sources of big data. As old scientific fields transform themselves, new ones emerge. Visual analytics entered our horizon in 2005 as a new field and has played a critical role ever since in advancing the science and technology for solving practical issues, especially when we deal situations that are full of complex, uncertain, incomplete, and potentially conflicting data. A representative case is concerned with maintaining the integrity of scientific literature itself. The increasing number of publications has overshadowed the increase of retractions. What can be done to maintain a trustworthy body of scientific knowledge?

What is the role that *Mapping Scientific Frontiers* has played? According to Google Scholar, it has been cited by 235 scientific publications on the web. These publications are in turn cited by an even broader range of articles. These articles allow us to take a glimpse on the context in which research in science mapping has been evolving. Interestingly, the citation profile appears to show two stages. The first one ranges from 2002 to 2008 and the second one from 2009 to the present (Fig. 1). Citations in the first stage peaked in 2007, whereas citations in the second stage were evenly distributed for the first 3 years. A study of citation data in the Web of Science revealed a similar pattern.

Fig. 1 The citation profile of *Mapping Scientific Frontiers* (Source: Google Scholar)

What is the citation pattern telling us? The nature of the set of articles that cited *Mapping Scientific Frontiers* as a whole can be analyzed in terms of how they are in turn cited by subsequently published articles. In particular, we turn to articles that have strong citation bursts, or abruptly increased citation rates, during the time span of 2002–2013. Figure 2 shows 25 articles of this type. Articles in the first stage shared a unique focus on information visualization and citation analysis. The original motivation of *Mapping Scientific Frontiers* was indeed to bridge together the two fields across the boundaries of different disciplines.

The second stage is predominated by a series of publications dedicated to global science maps at disciplinary levels as opposed to the document level in the first stage. The most influential work in the second stage in terms of citation burst is a 2009 *Scientometrics* article by Alan L. Porter and Ismael Rafolsonon on the interdisciplinarity of science. The second highest citation burst is attributed to a 2010 article published in the *Journal of American Society for Information Science and Technology* by Ismael Rafols, Alan L. Porter, and Loet Leydesdorff on science overlay maps. We are still in the second stage. In terms of the scale and the unit of analysis, the study of interdisciplinary interactions is a profound and potentially fruitful way to better understand the dynamics of scientific frontiers.

In addition to the conceptual and theoretical development, researchers today have a much wider range of choice than before in terms of computational tools for analyzing, visualizing, and exploring patterns and trends in scientific literature. Notable examples include CiteSpace, HistCite, VOSViewer, and Sci^2 for scientometric studies and science mapping; GeoTime, Jigsaw, and Tableau for visual analytics; and Gephi, Alluvial Maps, D3, and WebGL for more generic information visualization. Today, a critical mass is taking its shape and gathering its strengths as visual analytic tools, data sources, and exemplars of in-depth and longitudinal studies become increasingly accessible and inter-operable. Mapping Scientific Frontiers has reached a new level with a broad range of unprecedented opportunities to impact scientific activity across so many disciplines.

The second edition of *Mapping Scientific Frontiers* brings you some of the most profound discoveries and advances in the study of scientific knowledge and the dynamics of its evolution. Some of the new additions are highlighted as follows:

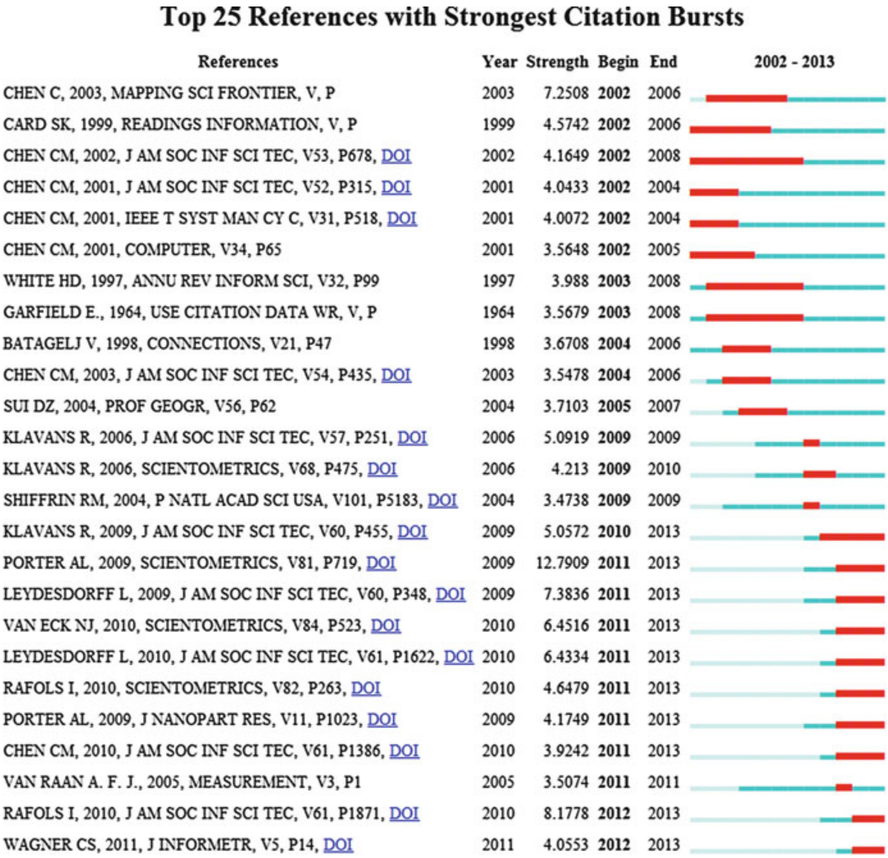

Fig. 2 A citation analysis of *Mapping Scientific Frontiers* reveals two stages of relevant research. *Red bars* indicate intervals of citation burst

- The Sloan Digital Sky Survey (SDSS) is featured in Chap. 2 in the context of how a map of the Universe may reveal.
- In Chap. 3, a series of new examples of visualizing a thematic evolution over time are illustrated, including the widely known ThemeRiver, the elegant TextFlow, and the versatile Alluvial Maps.
- Chapter 8 is a new chapter. It introduces the framework of a predictive analysis and demonstrates how it can apply to a fast-advancing field such as regenerative medicine, which highlights the work that was awarded the 2012 Nobel Prize in medicine on induced pluripotent stem cells (iPSCs). Chapter 8 also addresses practical implications of the retraction of a scientific publication. The second half of Chap. 8 is devoted to the design, construction, and analysis of global science maps, including our own new design of dual-map overlays.

- Chapter 9 is also a new chapter. It outlines some of the most representative visual analytic tools such as GeoTime and Jigsaw. It also describes major analytic features of CiteSpace.

The first edition concludes with ten challenges ahead. It is valuable to revisit these challenges identified over 10 years ago and see what have changed and what have newly emerged.

The second edition finishes with a set of new challenges and milestones ahead for mapping scientific frontiers.

15 April 2013 Chaomei Chen
Villanova
Pennsylvania, USA

Acknowledgements

The second edition in part reflects the result of a continuous research effort that I have been engaged in since the publication of the first edition. I'd like to acknowledge the support and contributions of my colleagues, students, and collaborators in various joint projects and publications, in particular, including my current and former students Timothy Schultz (Drexel University, USA), Jian Zhang (IBM Shanghai, China), and Donald A. Pellegrino (The Dow Chemical Company, USA), collaborators such as Pak Chung Wong (PNNL, USA), Michael S. Vogeley (Drexel University, USA), Alan MacEachren (Penn State University, USA), Jared Milbank (Pfizer, USA), Loet Leydesdorff (The Netherlands), Richard Klavans, Kevin Boyack, and Henry Small (SciTech Strategies, USA), and Hong Tseng (NIH, USA).

As a Chang Jiang Scholar, I have the opportunity to work with the WISELab at Dalian University of Technology, China, since 2008. I'd like to acknowledge the collaboration with Zeyuan Liu, Yue Chen, Zhigang Hu, and Shengbo Liu. Yue Chen is currently leading an ambitious effort to translate the second edition to Chinese.

I particularly appreciate the opportunity to work with Rod Miller, Chief Strategy Officer at the iSchool of Drexel, and Paul Dougherty, Licensing Manager at the Office of Technology Commercialization of Drexel University, over numerous fruitful discussions of various research topics.

I'd like to acknowledge the support of sponsored research and grants from the National Science Foundation (IIS-0612129, NSFDACS-10P1303, IIP 1160960), Department of Homeland Security, Pfizer, and IMS Health. I'd also like to take the opportunity to express my gratitude and appreciation to the hosts of my talks and keynote speeches, including Michael Dietrich, History of Biology, Woods Hole, MA; Stephanie Shipp, International Defense Agency (IDA), Washington D.C.; Paula Fearon, NIH, Bethesda, MD; David Chavalarias, Mining the Digital Traces of Science (MDTS), Paris, France; and Josiane Mothe, Institut de Recherche en Informatique de Toulouse, France.

Special thanks to Beverley Ford, Editorial Director Computer Science, Springer London Ltd, for her initiative and encouragement for getting the second edition on my agenda, and to Ben Bishop, Senior Editorial Assistant, Springer, for his clear and professional guidance that ensures a smooth and enjoyable process of preparation.

As always, my family, Baohuan, Calvin, and Steven, gives me all the strengths, the courage, and the inspirations.

Contents

Abbreviations

2dF	Two-degree field spectrograph
AAG	The Association of American Geographers
ACA	Author co-citation analysis
AGN	Active galactic nuclei
ANT	Actor Network Theory
ASE	Action Science Explorer
ASIS&T	American Society for Information Science and Technology
BSE	Bovine Spongiform Encephalopathy
CAS	Complex adaptive system
CBIR	Content-based image retrieval
CfA	Harvard-Smithsonian Center for Astrophysics
CJD	Creutzfeldt-Jakob Disease
C_{KL}	Centrality divergence
CL	Cluster linkage
DCA	Document co-citation analysis
DMO	Dual-map overlay
DNA	Deoxyribonucleic acid
DoD	Department of Defense
EOBT	Expert Opinion on Biological Therapy
GCS	Global Citation Score
GSA	Generalized Similarity Analysis
HDF	Hubble Deep Field
HMM	Hidden Markov Model
HUDF	Hubble Ultra Deep Field
iPSC	Induced pluripotent stem cell
KT Boundary	Cretaceous-Tertiary Boundary
LCS	Local Citation Score
LGL	Large Graph Layout
LLE	Locally linear embedding
LSI	Latent Semantic Indexing
MCR	Modularity change rate

MDS	Multidimensional scaling
MST	Minimum spanning tree
NB	Negative binomial
NVAC	National Visualization and Analytics Center
PCA	Principle component analysis
PFNET	Pathfinder network
PNNL	Pacific Northwest National Laboratory
PTSD	Post-traumatic stress disorder
SCI	Science Citation Index
SDSS	Sloan Digital Sky Survey
SOM	Self-organized maps
SSCI	Social Science Citation Index
SVD	Singular value decomposition
TRACES	Technology in Retrospect and Critical Events in Science
TREC	The Text Retrieval Conference
TSP	Traveling salesman problem
USPTO	The United States Patent and Trademark Office
WECC	The Western Power Grid
WoS	The Web of Science
XDF	eXtreme Deep Field
ZINB	Zero-inflated negative binomial

List of Figures