Chaomei Chen

Turning Points

The Nature of Creativity

高等教育出版社
HIGHER EDUCATION PRESS

Springer

Chaomei Chen

Turning Points

The Nature of Creativity

Chaomei Chen

Turning Points

The Nature of Creativity

With 82 figures, 18 of them in color

高等教育出版社
HIGHER EDUCATION PRESS

Springer

Author
Dr. Chaomei Chen
College of Information Science and Technology
Drexel University
3141 Chestnut Street, Philadelphia
PA 19104-2875, USA
E-mail: chaomei.chen@drexel.edu

ISBN 978-7-04-031703-9
Higher Education Press, Beijing

ISBN 978-3-642-19159-6 e-ISBN 978-3-642-19160-2
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011920985

Printed on acid-free paper

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

Foreword

Among the uniquely human capabilities is the capacity to create and discover. Understanding how humans create innovative art, music, poetry, or novels and discover scientific principles patterns, or relationships requires a recursive form of creativity and discovery.

The foundations for human creativity and discovery depend on passion for solving problems and fluency with social contexts that promote solutions. The passion produces persistence over time and enables devotion to solving important problems, filling troubling gaps, stretching annoying boundaries, or opening doors to fresh opportunities.

The fluency with social contexts helps researchers to see problems more clearly, bridge disciplines, and apply methods from one knowledge domain to another. The social context also provides powerful motivations that encourage varied forms of competition and collaboration. Sometimes competition is fierce, other times it can be friendly. Sometimes collaboration is narrow and limited to dialogs between trusted partners, other times it can be broad and long-term, producing lively conversations among thousands of contributors who are united by the passion to solve a problem. Innovators who protect their nascent ideas too closely will miss the opportunity to get feedback about their progress or learn about related ideas.

Researchers are increasingly attracted to study the dynamics of creativity and discovery. For the first time in history the databases of human scientific activity are sufficiently large and widely available. For the first time in history the tools for analyzing this data are capable of performing appropriate analyses and becoming widely available.

Retrospective citation analysis of scientific papers remains the major approach, sometimes complemented by informed ethnographic observations and interviews by researchers with sufficient knowledge-domain understanding to recognize important steps, controversies, or mistakes. However, analysis of patents, patent citations, trade journal articles, blogs, emails, twitter posts, and other social media will provide a finer-grained, more diverse, and more immediate record of how scientific breakthroughs emerge.

Citation analysis goes far beyond simple counts of who cited whom, but expands to author co-citation and document co-citation networks, while adding potent metrics such as betweenness centrality to find boundary-spanning papers that bridge knowledge domains. An important tool for these analyses is network visualization, which sometimes surprises researchers by showing important clusters, revealing bridging papers, or spotting important papers that may be tragically ignored for many years or become very hot quickly.

This latest book from Chaomei Chen makes important contributions to research on creativity because he brings a remarkably broad perspective to this topic, weaving together several strands of research. Chen clarifies existing theories, applies interesting metrics, and shows compelling visualizations. He lets readers know exactly what his point of view is: "transformative discoveries are likely to emerge from the twilight zones where multiple fields meet." This strong conviction is validated by retrospective analyses and case studies from impressively diverse branches of science.

The importance of this book, *Turning Points The Nature of Creativity*, is that Chen has a greater ambition than to look back, he wants to be in the moment by offering researchers the capacity to see what is currently happening in their knowledge domains, so as to spot important contributions early. The capacity to predict which papers will eventually be highly cited would be a wonderful gift to researchers, government policy planners, and industry managers. This goal is not easy to attain, but Chen suggest some promising possibilities.

The even more ambitious challenge that Chen takes on is to spot opportunities for interesting research by identifying "structural holes" or missing intersections of related knowledge domains. This is not easy since there are many unproductive intersections, so it takes informed expertise to make the right judgments or spot early signs of progress. This is a seductive idea, but Chen warns of many forms of "biases, pitfalls, and cognitive traps." Still he boldly offers a powerful claim: "a paper with a high betweenness centrality is potentially a transformative discovery. In addition, it would be possible to use this metric to identify potential future discoveries by calculating the would-be betweenness centrality of a hypothetical connection between two disparate areas of existing knowledge networks....Thus, betweenness centrality can be translated into interestingness, which can be in turn translated into actionability."

Readers should take time to reflect on the goals Chen lays out and appreciate the diverse sources he draws from. They should also carefully consider the metrics he proposes and study the visualizations from his CiteSpace system. Chen admirably lays out his emerging ideas, seeking constructive dialogs and engaging in fruitful conversations. This makes for provocative reading and stimulates fresh thinking. Readers can respond with even better theories, data, metrics, and visualization.

Ben Shneiderman
University of Maryland
July 2011

Preface

Research assessment has become a central issue for more and more government agencies and private organizations in making decisions and policies. New indicators of research excellence or predictors of impact are popping out one after another. However, if we look behind the available methods and beyond the horizon decorated by the various types of indicators, then we will encounter a few questions again and again: What is the nature of creativity in science? Is there a way that we can tell great ideas early on? Are there ways that can help us to choose the right paths? Can we make ourselves more creative?

There are only two types of theories no matter what their subjects are: the ones that are instructional and the ones that are not. An instructional theory will explain the underlying mechanisms of a phenomenon in such a way that we can see what we need to do to make a difference. The quest for us in this book is to look for a better understanding of mechanisms behind creativity, especially in the context of making and assessing scientific discoveries. In this book, my goal is to identify principles that appear to be necessary for creative thinking from a diverse range of sources and clarify where we may struggle with biases and pitfalls created by our own perceptual and cognitive systems. Then I will introduce an explanatory and computational theory of discovery and demonstrate its instructional nature through a series of increasingly refined quantitative approaches to the study of knowledge domains in science. Finally, the potential of transformative research is measured by metrics derived from the theoretical underpinning and validated with retrospective indicators of impact. The theory, for example, leads to a much simplified explanation of why some of the good predictors of citation counts of an article found by previous research are due to the same underlying mechanisms.

The conception of the theory of discovery was inspired by a series of intellectual landmarks across a diverse range of perspectives, notably, Vannevar Bush's *As We May Think* and his vision for trailblazing a space of knowledge in his Memex (memory and index), Thomas Kuhn's paradigm shift theory of scientific revolutions, Henry Small's methods for analyzing co-citation networks, Ronald Burt's structural-hole theory, and Peter Pirolli's optimal information foraging theory. The development and use of the CiteSpace system have played an instrumental role in experimenting and synthesizing these great ideas. I have been developing and maintaining CiteSpace since 2003. I have made it freely available for researchers and students to analyze emerging trends and turning points in the literature. The provision of CiteSpace has probably also promoted the awareness of scientometrics, the field that is concerned with quantitative approaches to the study of science. Feedback, questions, and requests for new features from a diverse and growing population of users have also propelled the search for theories to explain various patterns that we see in the literature.

The central thesis of the book is that there are generic mechanisms for creative thinking and problem solving. If we can better understand these mechanisms, then we will be able to incorporate them and further enhance them with computational techniques. Another important insight gained from reviewing the literature across different fields is that creativity is about the ability and willingness to find a new perspective so that we can see something that we take for granted.

The notion of an intellectual turning point has naturally emerged. Kuhn's gestalt switch between competing paradigms and Hegel's syntheses of theses and antitheses are exemplars of view-changing intellectual turning points. We may feel lucky or unlucky, depending on the particular perspective we take. We may miss the obvious if we are looking for something else. I hope that this book can provide the reader with some useful perspectives to study science and its role in society as well as insights into the nature of creativity so that we will be better able to recognize creative ideas and create opportunities for more creative ideas.

I have a few types of readers in mind when I was preparing for this book:

1) anyone who is curious about the nature of creativity and wondering if there is anything beyond the serendipitous view of creativity
2) analysts, evaluators, and policy makers in a situation where tough decisions have to be made that will influence the fate of creative work
3) researchers and students who need to not only keep abreast of their own fields of study but also position themselves strategically with a competitive edge
4) historians and philosophers of science

The first four chapters of the book should be accessible to college students and more advanced levels. The next four chapters may require a higher level of background information in areas such as network analysis and citation analysis. The book may be used for graduate-level courses or seminars in information science, research evaluation, and business management.

Chaomei Chen
Philadelphia, Pennsylvania
April 2011

Acknowledgements

Many people have played an information role in the ideas presented in this book.

My long-term collaborators in interdisciplinary research projects include Michael S. Vogeley, an astrophysicist at the Department of Physics, Drexel University, on a project funded by the National Science Foundation (NSF) (IIS-0612129) to study the interconnections between astronomical literature and the usage of the astronomical data obtained by the Sloan Digital Sky Survey (SDSS), Alan M. MacEachren, at the Department of Geography, Penn State University, on the Northeast Visual Analytic Center (NEVAC) project funded by the Department of Homeland Security, my graduate research assistants and doctoral students Jian Zhang and Don Pellegrino, and international visitors Fidelia Ibekwe-SanJuan (France) and Roberto Pinho (Brazil).

Eugene Garfield and Henry Small, visionary pioneers of citation analysis and co-citation analysis at Thomson Reuters, have been generous with their time and insights. Thomson Reuters' younger generation, David Liu (China), Weiping Yue (China), and Berenika Webster (Australia), are enthusiastic, energetic, and supportive. In particular, Thomson Reuters made generously arrangements for me to have an extensive period of access to the Web of Science while I was on sabbatical leave. I was a recipient of the 2002 Citation Research Award from the ISI and the American Society for Information Science and Technology.

I would like to thank Julia I. Lane and Mary L. Maher, Program Directors at the National Science Foundation (NSF), for their masterminded efforts in organizing the research portfolio evaluation project to explore technical feasibilities of evaluating NSF proposals (NSFDACS-10P1303), Jared Milbank and Bruce A. Lefker at Pfizer Global Research and Development at Groton Labs for collaborating on a Pfizer-funded drug discovery project.

I am also grateful to Zeyuan Liu at the WISELab, Dalian University of Technology, for his enthusiasm, vision, and insights in the use of CiteSpace in mapping knowledge domains in China, Hung Tseng, a biologist-turned NIH program director, for sharing his enthusiasm and insights in issues concerning the evaluation of research and tracing timelines of discoveries from a funding agency's point of view, Rod Miller, Drexel University, for numerous in-depth conversations on my current research and on articulating and communicating complex ideas effectively, and Ying Liu, the editor at the Higher Education Press, China, for her initiative and efforts in getting the book writing project underway.

To Baohuan, Calvin, and Steven, my caring, loving, and cheerful buddies in my sweet family, thank you for everything.

Contents

Chapter 1 The Gathering Storm

There are two ways to boil a frog alive. One is to boil the water first and then drop the frog into boiling water — the frog will jump out from the immediate crisis. The other is to put the frog in cold water and then gradually heat the water until it boils — the frog will not realize that it is now in a creeping crisis. As far as the frog is concerned, the creeping crisis is even more dangerous because the frog loses its chance to make a move that could save its life.

Several major crises in the past triggered the U.S. to respond immediately, notably the Japanese attack at Pearl Harbor in 1941, the Soviet Union's launch of Sputnik 1 in 1957, and the 911 terrorist attacks in 2001. The Sputnik crisis, for example, led to the creation of NASA and DARPA and an increase in the U.S. government spending on scientific research and education. In contrast to these abrupt crises, several prestigious committees and advisory boards to the governing bodies of science and technology policy have sounded an alarm that the U.S. is now facing an invisible but deeply profound crisis — a creeping crisis that is eroding the very foundation that has sustained the competitive position of the nation in science and technology.

In 2005, William Wulf, the President of the National Academy of Engineering (NAE), made his case before the U.S. House of Representatives' Commission on Science. He used the creeping crisis scenario to stress the nature of the current crisis — a pattern of short-term thinking and a lack of long-term investment. However, the view is controversial. There have been intensive debates on the priorities that the nation should act upon and whether there is such a thing as a "creeping crisis" altogether. One of the central points in the debate is whether the science and engineering (S&E) education, especially math and science, is trailing behind the major competitors in the world in terms of standard test performance and the ability to meet the demand of the industries.

Why are people's views so different that the idea of any reconciliation seems to be distant and far-fetched? Is the crisis really there? Why are some so concerned while others not? What are the key arguments and counterarguments? After all, what I want to address in this book is: what are the most critical factors that hinge the nation's leading position in science and technology? Furthermore, what does it really take to sustain the competitiveness

C. Chen, *Turning Points*

of the U.S. in science and technology?

1.1 The Gathering Storm

The notion that the U.S. is in the middle of a creeping crisis was most forcefully presented to the U.S. House of Representatives' Committee on Science on October 20, 2005[1]. Norman R. Augustine, the chairman of the competitiveness assessment committee, P. Roy Vagelos, a member of the committee, and William A. Wulf, the president of the National Academy of Engineering presented their assessments of the situation. Augustine is the retired chairman and CEO of Lockheed Martin Corporation and Vagelos is the retired chairman and CEO of Merck. The full report was published by the National Academies Press in 2007, entitled *Rising above the Gathering Storm* (National Academy of Sciences, National Academy of Engineering, & Institute of Medicine of the National Academies, 2007). In the same year, *Is America Falling Off the Flat Earth?*, written by Augustine, was also published by the National Academies Press[2] (Augustine, 2007).

The Gathering Storm committee included members such as Nobel laureate Joshua Lederberg, executives of research-intensive corporations such as Intel and DuPont, the director of Lawrence Berkeley National Laboratory, and presidents of MIT, Yale University, Texas A&M, Rensselaer Polytechnic Institute, and the University of Maryland. The prestigious background of the committee and its starry members as well as the well articulated arguments have brought a considerable publicity to the notion of the creeping crisis — the gathering storm!

The key points of the creeping crisis presented in the Gathering Storm committee can be summarized as follows:

1) America must repair its failing K-12 educational system, particularly in mathematics and science.
2) The federal government must markedly increase its investment in basic research, that is, in the creation of new knowledge.

The primary factor in this crisis is the so-called the Death of Distance, which refers to the increasing globalization in all aspects of our life. Now the competitors and consumers are all just a "mouse-click" away. Fast and profound changes in a wide range of areas are threatening the leading position of the U.S., for example, the mobility of manufacturing driven by the cost of labor and the existence of a vibrant domestic market. For the cost of one engineer in the United States, a company can hire eleven in India. More importantly, the Gathering Storm committee highlighted that the increasing mobility of financial capital, human capital, and knowledge capital is now

[1] http://www7.nationalacademies.org/ocga/testimony/gathering_storm_energizing_and_employing_america2.asp

[2] The National Academies Press offers a free podcast free of charge at http://books.nap.edu/catalog.php?record_id=12021

accelerating and deepening the crisis. On the other hand, competitors in other countries have recognized the key mechanisms that sustain America's competitiveness and are seeking to emulate the best of the America's system. To assure that the U.S. does not fall behind the race, there is clearly a sense of urgency. According to Augustine,

> *It is the unanimous view of our committee that America today faces a serious and intensifying challenge with regard to its future competitiveness and standard of living. Further, we appear to be on a losing path. We are here today hoping both to elevate the nation's awareness of this developing situation and to propose constructive solutions.*

Charles Darwin observed that "it is not the strongest of the species that survives, nor the most intelligent, but the one most responsive to change." In 1993, the Committee on Science, Engineering, and Public Policy (COSEPUP) recommended that the United States needs to be among the world leaders in all fields of research in order to sustain the following key abilities:

- Bring the best available knowledge to bear on problems related to national objectives even if that knowledge appears unexpectedly in a field not traditionally linked to that objective.
- Quickly recognize, extend, and use important research results that occur elsewhere.
- Prepare students in American colleges and universities to become leaders themselves and to extend and apply the frontiers of knowledge.
- Attract the brightest young students.

The Gathering Storm committee has made a compelling case of a profound sense of urgency and the need for action. The array of evidence include the choice of investment: in 2005, for the first time in 20 years, U.S. investors put more new money into international stock funds than into U.S. stock funds. The overseas fraction of newly invested stock funds in the U.S. changed from 8% in 1999 to 77% in 2005. In a survey of the attractive locations for new R&D facilities, 41% of the global corporations voted for the U.S. and 62% for China. Augustine quoted a poem by Richard Hodgetts to sum up the urgency of the serious and intensifying challenge to America's future competitiveness and standard of living in a global environment:

> *Every morning in Africa a gazelle wakes up.*
> *It knows it must outrun the fastest lion or it will be killed.*
> *Every morning in Africa a lion wakes up.*
> *It knows it must outrun the slowest gazelle or it will starve.*
> *It doesn't matter whether you're a lion or a gazelle —*
> *when the sun comes up, you'd better be running.*

Augustine (2007) noted that he was astonished by the degree to which foreign officials are familiar with the Gathering Storm report. The Doomsday Scenario, as he described, would be the Gathering Storm succeeded in motivating others to do more and then the U.S. did or sustained little. The

U.S. Congress has passed the America COMPETES Act[3] in 2007 to enact some of the recommendations made by the Gathering Storm committee. For example, the Act includes requirements to the National Science Foundation (NSF), the major funding agency of basic research:

- (Sec. 4006) Requires the NSF Director to: (1) consider the degree to which NSF-eligible awards and research activities may assist in meeting critical national needs in innovation, competitiveness, the physical and natural sciences, technology, engineering, and mathematics; and (2) give priority in the selection of the NSF awards, research resources, and grants to entities that can be expected to make contributions in such fields.
- (Sec. 4007) Prohibits anything in Divisions A or D of this Act from being construed to alter or modify the NSF merit-review system or peer-review process.
- (Sec. 4008) Earmarks funds for FY2008-FY2011 for the Experimental Program to Stimulate Competitive Research under the National Science Foundation Authorization Act of 1988.

Despite the compelling creeping crisis case and the consensus of the need for action, many have raised serious questions that challenge the diagnostics and treatments of the crisis. Indeed, multiple views, conflicting positions, and competing recommendations need to be validated, resolved, and implemented. Not only for policy makers but also for scientists, educators, students, and the general public, there is the urgent need for making sense of what is really happening, and more importantly for understanding the spectrum of the long-term consequences of decisions made today.

1.2 Into the Eye of the Storm

One of the most forceful attacks of the Gathering Storm report is made by *Into the Eye of the Storm* (Lowell & Salzman, 2007). The authors of the paper are Lindsay Lowell of Georgetown University and Hal Salzman of the Urban Institute. Their research was funded by the Alfred P. Sloan Foundation and the National Science Foundation.

The key finding of the *Into the Eye of the Storm* is that their review of the data fails to find support for the challenges identify by the Gathering Storm and those with similar views. Specifically, they did not find evidence for the decline in the supply of high quality students from the beginning to the end of the science and engineering pipeline due to a declining emphasis on mathematics and science education and a declining career interest among the U.S. domestic students in science and engineering careers. First, Lowell and Salzman showed that the claim that the U.S. falls behind the world in science and mathematics is questionable; their data shows that the U.S. is the only country with a considerable diversity of student performance and

[3] http://thomas.loc.gov/cgi-bin/bdquery/z?d110:SN00761:@@@D&summ2=m&

that simple rank positions make little sense in light of such a degree of diversity. Second, their analysis of the flow of students up through the science and engineering pipeline suggests that the supply of qualified graduates is far in excess of demand. Third, the more than adequate supply requires a better understanding why the demand side fails to induce more graduates into the S&E workforce. Policy approaches to human capital development and employment from the prior era do not address the current workforce or economic policy needs.

Lowell and Salzman's analysis shows that, from employers' point of view, literacy and a competence in a broad range of subjects beyond math and science are essential. Furthermore, they rightly stated that the question is not about whether to improve the U.S. education system, but rather why the U.S. performance is lower than other countries, what the implications are for the future competitiveness, and what polices would best address the deficiencies. Their analysis draws attention to the fact that, according to the 2006 U.S. census, single-parent households with children under age 17 account for 33% of families in the U.S., whereas the number is 17% in Norway and less than 10% in Japan, Singapore, and Korea. Therefore, it is unclear whether using average test scores provide any meaningful indication of education or potential economic performance of the U.S. because one could argue that it is the diversity and openness of the U.S. that contribute to its lower average educational performance as well as its high economic performance.

Further analysis of the education-to-career pipeline shows that science and engineering firms most often complain about schools failing to provide students with the non-technical skills needed in today's firms.

In summary, *Into the Eye of the Storm* concluded that the perceived labor market shortage of scientists and engineers and the decline of qualified students are not supported by the educational performance and employment data that Lowell and Salzman have reviewed. In contrast to the policy focus of the U.S. competitiveness committees calling for the U.S. to emulate Singapore's math and science education programs, Singapore's recent competitiveness policy focuses on creativity and developing a more broad-based education — an emulation of the U.S. education.

The debates have made it clear that different questions should be asked: What are the factors that have led to the consistent high performance of the U.S. economy? What kind of workforce is likely to improve prospects of the U.S. in the future? Lessons learned from the conflicting views underline that evidence-based policy is necessary for developing effective programs for the emerging global economy. Julia Lane, the Program Director of the NSF Science of Science Policy Program, supports evidence-based approaches to science policy.

In a recent article published in the *Scientific American*, Beryl Lieff Benderly (2010), a columnist for the Science Careers of the journal *Science*, addressed the question: Does the U.S. produce too many scientists? For example, she addressed practical issues associated with the fact that labs in the

U.S. are typically staffed by graduate students and postdoctoral researchers and new generations of graduates face an increasingly tough situation to land on a tenure track position in universities in the U.S.. Her article quickly attracted over 200 comments within days. Most comments spoke for personal experience in moving up along the education-to-career pipeline that the *Into the Eye of the Storm* studied.

A manager in an engineering organization commented on what skills are needed in his/her organization:

> "*As a manger in an engineering organization, what I need are talented BS and MS level engineers interested in hardware design, not PhD researchers interested in basic science. Innovation that brings items to market drives the economy, not fundamental research. Only when the economy is producing marketable products can we afford the luxury of basic science; not to belittle the importance of science, but its rewards are less immediate.*"

In terms of the metaphor of the education-to-career pipeline, BS and MS level engineers would leave the pipeline much earlier than those who graduate with their PhDs.

In contrast, another commentator addressed the range of career options concerning the far end of the pipeline, i.e. graduates with a Ph.D. and challenged the notion that the best career move for a Ph.D. is a tenure track position in a research university:

> "*A very valid and fruitful path is high-level engineering and science in the industrial sector. I dare say that Google and Microsoft have more PhDs than many universities.*"

Yet another commentator expressed a similar view:

> "*When people say that we need more scientists and engineers, this is not what they mean. What they are talking about is the need for more scientists and engineers that are going to tackle the difficult problems of our time and being required to regularly justify what research you want (academic or industrial) funding is a good thing.*"

Although the more attractive paycheck elsewhere in professions such as finance and law is often used to explain why people abandon the science and engineering career pipeline, some has expressed the view that scientists should not be wealth seekers. For example, an european reader made the following comments:

> "*As scientists, we don't (or shouldn't) pursue wealth. We do appreciate a decent salary, AND more important (ly), some guarantee of stability, a pension and a decent health care coverage ...*"

A different reader pointed out that the reliance of America's science on immigrants is not something new:

> "*American science has always had a strong representation by immigrants. We just need to go through the list of Nobel laure-*

> *ates for example. So there is nothing like 'Reversing the trend' to how it was. It's always been like that. The authors constant reference to "Native born white men" is inappropriate. A large number of American high school students excelling in international math and science competitions as pointed by the author himself are ethnically Asian or from India. Think about Steven Chu for example.*"

Throughout the widespread debates, some argue that the current volatile research system is no more than a source of instability, and it is the instability that drives many graduates off their originally chosen career paths. On the other hand, others believe that the instability and constant competition is precisely where the U.S. science and technology draws its competitive strength. This is the same kind of natural selection Charles Darwin talked about: the fittest will survive!

1.3 The Yuasa Phenomenon

A phenomenon first identified in 1960s by Japanese historian and physicist Mintomo Yuasa (1909 – 2005) may provide a different perspective to the already heated debate over the Gathering Storm. Yuasa analyzed world scientific activity records compiled from *the Chronological Table of Science & Technology* (in Japanese) and Webster's *Biographical Dictionary of Names of Noteworthy Persons with Pronunciations and Concise Biographies*. He studied the trajectories of countries that claimed more than 25% of the major scientific achievements of the entire world in the history and defined such countries as the centers of scientific activity. He noticed that the center of scientific activity appears to move from one country to another periodically, every 80∼100 years! This is the Yuasa Phenomenon (Yuasa, 1962).

Italy was the center for 70 years from 1540 till 1610. England was the center for 70 years from 1660 till 1730. France was the center for 60 years from 1770 till 1830. Germany was the center for 110 years (1810 – 1920). The most interesting one is the current center — the U.S.. The U.S. became the current center 90 years ago, since 1920. According to the periodical pattern found by Yuasa, the shift of the U.S. as the world scientific activity center could take place between 2000 and 2020. An equally profound question is: if the center does move, which country is likely to be the next? If the Gathering Storm debate is viewed in this context, one has to wonder about the reasons behind such shifts.

What can cause the center to drift away? Or equivalently, what makes the center to stay? Chinese scholar Hongzhou Zhao, not aware of Yuasa's work, independently discovered the same phenomenon. Zhao's work was introduced to the Western in 1985 (Zhao & Jiang, 1985). However, it seems that their work is still not widely known to the western world — as of 2010, their paper has been cited three times. It was first cited in 1987 by Schubert (1987) in a *Scientometrics* article on quantitative studies of science. In 1993, it was cited in *Psychological Inquiry* by Hans Eysenck (1993) on creativity and personality. He suggested a causal chain reaching from DNA to creative achievement, based largely on experimental findings not usually considered in relation to creativity (e.g., latent inhibition). His model is highly speculative, but nonetheless testable. The most recent citation to Zhao and Jiang's paper was made by an article on a bibliometric model for journal discarding policy in academic libraries.

Zhao introduced the notion of the social mean age of a country's scientists at time t as the average age of a scientist makes significant contributions:

$$A_t = \sum_{i=1,\ldots,n} \frac{X_i - X_b}{N_t}$$

where X_b is the year of the birth of a scientist, X_i is the time when the scientist makes noteworthy contributions, and N_t is the total number of scientists at time t. Zhao noticed some interesting patterns: the A_t of 50 years old seems to be a tipping point. Immediately before a country becomes the center of scientific activity, the A_t of its outstanding scientists is below 50 years old. For example, Italy was the world center in 1540 – 1610; the social mean age of scientists of Italy was 30∼45 years old between 1530 and 1570. Similarly, England was the center during 1660 – 1730 and its social mean age was 38∼45 between 1640 and 1680. France was the center 1770 – 1830 and its social mean age was 43∼50 between 1760 and 1800. Germany became the center in 1810 – 1920 and its social mean age was 41∼45. The U.S. has been the center since 1920 and its social mean age of scientists was about 50 between 1860 and 1920.

On the other hand, if the social mean age of scientists in the host country of the current center of scientific activity exceeds 50 years old, it tends to lose its center position. For example, the A_t of France started to exceed 50 years old in 1800; by 1840, the center shifted to England. Why is the age of 50 so special?

As we shall see in Chapter 2, Zhao approached to this question from a statistical perspective and defined the concept of an optimal age — a period of the most creative years in the career of a scientist. Zhao found that when a country's social mean age approaches the distribution of the optimal ages of the scientists in the country, the country's science is likely on the rise; otherwise, it is likely to decline. The estimation of the optimal age is built on his theory of scientific discovery. We will re-visit Zhao's work in more detail in Chapter 2.

A different approach to the question was offered by Zeyuan Liu and Haishan Wang in 1980s[4]. They found that a country's status of the world center of scientific activities appeared to follow a 60-year leading period of revolutions

[4] http://www.collnet.de/workshop/liu.html

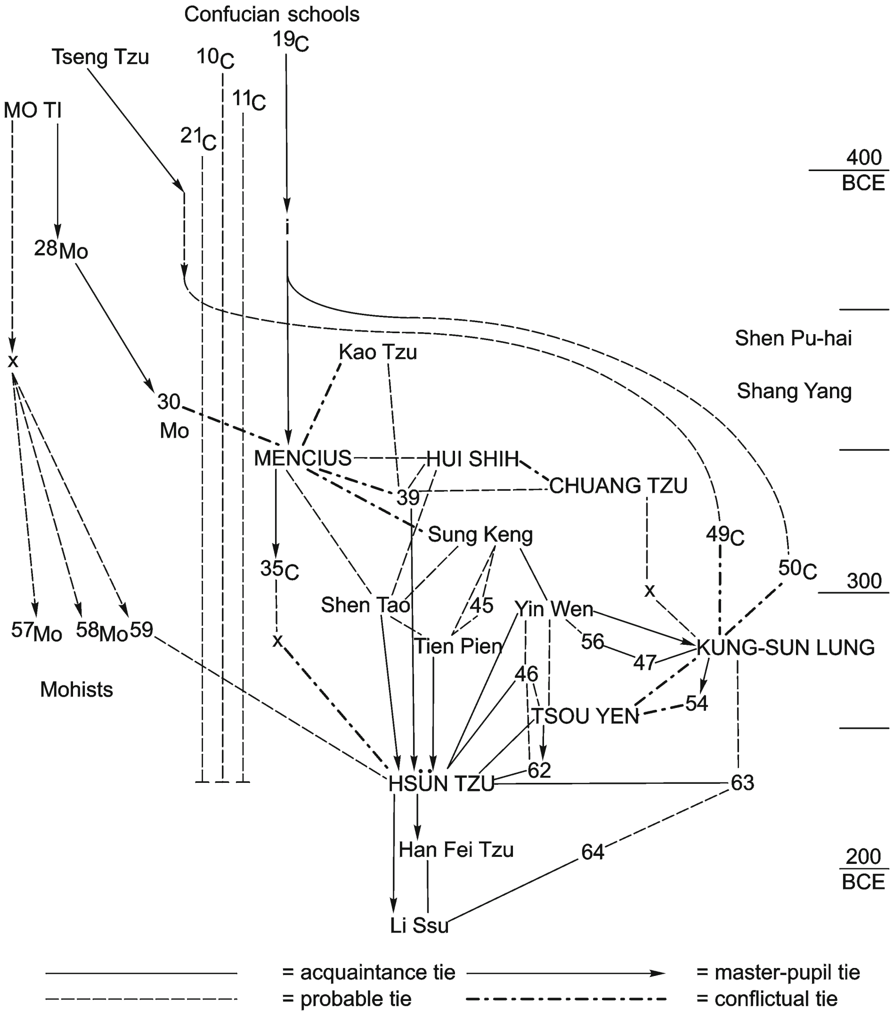
of philosophy in the same country. In other words, philosophical revolutions lead scientific revolutions. Furthermore, a macroscopic chain of revolutions was found in England, France, and Germany: philosophical → political → scientific → industrial revolutions. For example, Italy experienced its philosophical revolution in 1480, which was 60 years before it became the scientific center of the world in 1540. England's philosophical revolution began in 1600, also 60 years ahead of its status as the world center of scientific activities in 1660.

The social mean age of scientists, the optimal age of scientists in a country, and the presence or absence of a philosophical revolution provide a set of interesting macroscopic-level indicators. On the other hand, finer-grained theories and models of scientific discovery are necessary to investigate any substantial connections underlying these observations. Furthermore, while macroscopic observations provide interesting backdrops of scientific activities, many questions are unlikely to be answered precisely unless we take the development of scientific fields into account.

1.4 Transformative Research and the Nature of Creativity

The Death of Distance is ubiquitously behind the globalized and intensified competitions in and across all areas of economy, culture, politics, education, and science and technology. Taxpayers, small business, large corporate companies, schools and universities, and government agencies are all under tremendous pressure to act. Darwin's natural selection is undertaking a whole new wave of variations and taking place at an unprecedented rate and scale.

From a sociological perspective of the philosophy of science, Randall Collins (1998) argued that intellectual life is first of all conflict and disagreement. His insight is that the advance of an intellectual field is very much due to rivalry and competing schools of thought that are often active within the same generational span of approximately 35 years. He introduced the notion of attention space and argued that "creativity is the friction of the attention space at the moments when the structural blocks are grinding against each other the hardest.". The attention space is restructured by pressing in opposing directions. He spent over 25 years to assemble intellectual networks of social links among philosophers whose ideas have been passed along in later generations. He constructed such networks for China, India, Japan, Greece, modern Europe, and other areas over very long periods of time. He used a generation of philosophers as a minimal unit for structural change in an intellectual attention space. For example, it took 6 generations to move from Confucius to Mencius and Chuang Tzu along the Chinese intellectual chains. Fig. 1.1 shows an example of the intellectual network of Chinese philosophers between 400B.C. and 200B.C.. A major difference between Collins' grinding attention space and Kuhn's competing paradigms is that for Collins explicit rivalry between schools of thought often developed in succeeding generations



ALL CAPS (e.g., CHUANG TZU) = major philosopher
Lowercase (e.g., Li Ssu) = secondary philosopher
number = minor philosopher, listed in key (see Appendix 3)
(e.g., 10 is Tzu-yu, founder of school of Confucian disciples)
i = incidental person (not known independently of contact with a known philosopher)
x = unnamed person
C = Confucian school
Mo = Mohist school

Fig. 1.1 A social-intellectual network of Chinese philosophers (400 – 200 B.C.). Source: Figure 2.1 in (Collins, 1998, p. 55).

(Collins, 1998.), whereas Kuhn's competing paradigms are simultaneous. Major philosophers (labeled with all capital letters) such as Mencius and Kung-Sun Lung are all at the center of colliding perspectives.

Fig. 1.2 depicts the intellectual footprints of the field of nanoscience between 1997 and 2007. It shows how fast a field has been moving forward and how much of the literature has been left behind. Each small dot in the image represents an article that was cited by researchers in the field. Each dot is surrounded by tree-rings of citations that the article received. A bigger-sized disc indicates that the corresponding article has been cited more often than an article with a smaller-sized disc. The majority of the articles on the left-hand side were cited by the field at the beginning of the timeframe, i.e. late 1990s. The bluish colors of articles in this area indicate that the field no longer cited much of them for a long time. In contrast, the right-hand side is full of recent activities. The colors of the citation rings in this area are warmer and brighter, indicating more recent citations. The citation tree rings of a few articles have layers of rings in red. It means that these articles experienced a significant surge of citations. They were at the center of the attention of the field. They were the hot topics.

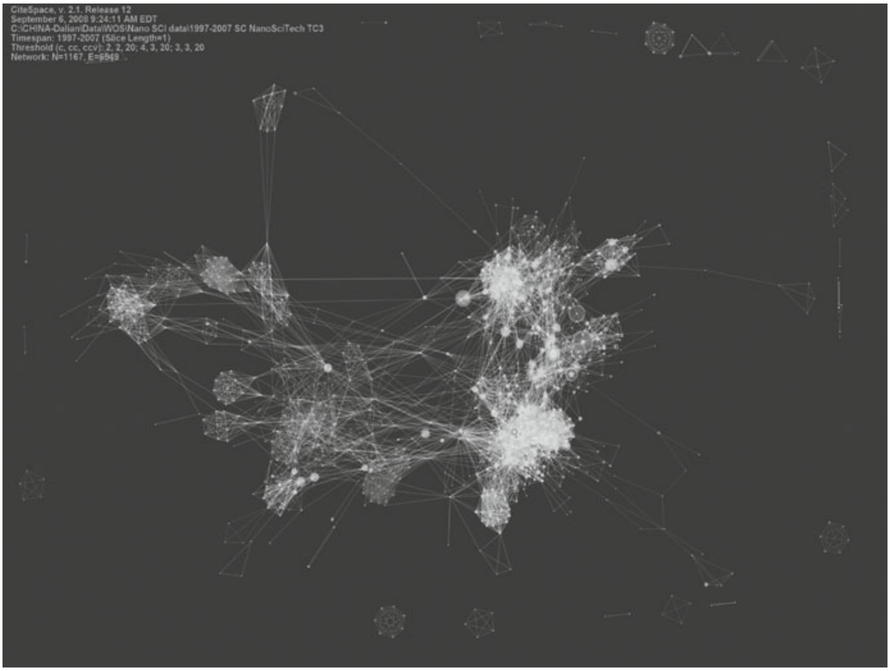

CiteSpace, v. 2.1. Release 12
September 6, 2008 9:24:11 AM EDT
C:\CHINA-Dalian\Data\WOS\Nano SCI data\1997-2007 SC NanoSciTech TC3
Timespan: 1997-2007 (Slice Length=1)
Threshold (c, cc, ccv): 2, 2, 20; 4, 3, 20; 3, 3, 20
Network: N=1167, E=6549

Fig. 1.2 The intellectual trails of the field of nanoscience between 1997 and 2007. (see color figure at the end of this book)

Fig. 1.3 shows not only a map of the Universe but also discoveries and research interests associated with various areas in the Universe. The earth is at the center of the map because the distance to an astronomic object is measured from the earth. The blue band of galaxies and the red band of quasars were formed at the early stage of the Universe. As the Universe expands, they become further away from us. The Hubble Ultra Deep Field, shown at the upper-right corner of the image, was one of the farthest observations made by scientists. Unlike the free-form layout method used in generating the visualization shown in Figure 1.2, the map of the Universe preserves the relative positions of astronomic objects. It is common in cartography to use a base map as the general organizational framework and then add various thematic layers on top of it. Adding multiple thematic layers is in effect combining information from multiple perspectives. A fundamental question yet to be answered is how one should interpret the meaning of such combinations. Each perspective represents its own conceptual space, which may or may not be compatible with other spaces. The compatibility here means whether there exists a topological mapping from one space to another. A central property of topological mapping is that it reserves the proximity relations so that nearby points in one space will remain to be neighbors when they are mapped into a new space. This is obviously not held between the astronomical space and the space of astronomical knowledge. Two black holes may be further apart in the Universe, but they can be dealt with by the same theory in the knowledge space. In contrast, two different theories may address the same phenomenon in the Universe.

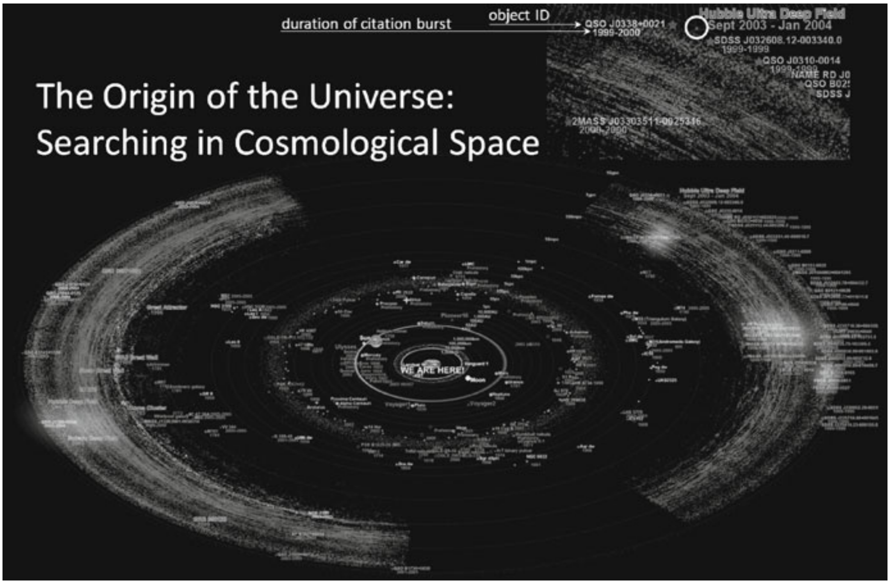


Fig. 1.3 A map of the Universe with overlays of discoveries and astronomical objects associated with bursts of citations. The close-up view of the Hubble Ultra Deep Field is shown at the upper-right corner (circled). (see color figure at the end of this book)

In May 2009, as H1N1 was rapidly spreading across many countries, there was a rich body of knowledge about influenza pandemics in the literature. The Web of Science alone indexed over 4,500 research papers on influenza and pandemics. Fig. 1.4 shows a timeline visualization of this literature as of May 8th, 2009.[5] Spots in red were articles with a burst of citations. In contrast, Fig. 1.5 shows a similarity map of 114,996 influenza virus protein sequences. Some of the significant questions to be addressed are what multiple views of influenza such as these two would tell us and how they would foster new research questions.

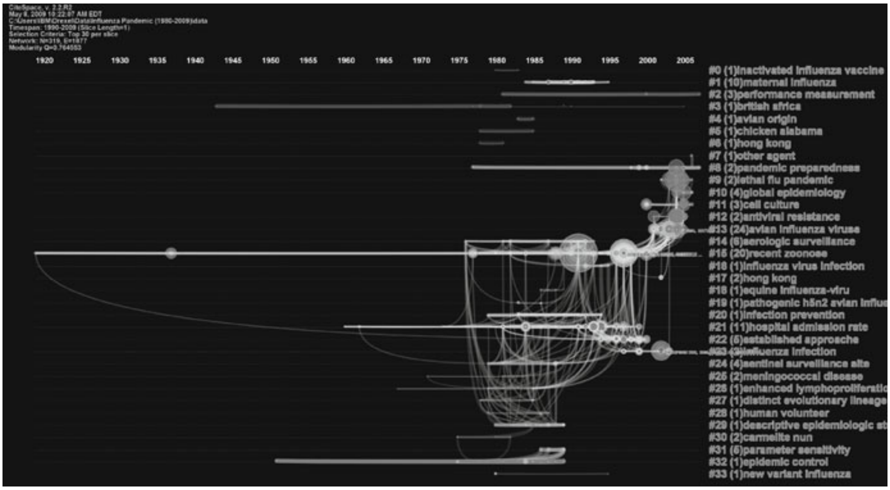


Fig. 1.4 A timeline visualization of the state of the art in research related to influenza and pandemics as of May 8th, 2009. (see color figure at the end of this book)

As we can see, this type of mismatch between multiple conceptualizations can be also found in many other disciplines, for example, chemical space versus biological space in drug discovery and world views from competing paradigms. The conflicts of conceptualizations present the potential of discovery.

There is a growing and widening interest in searching for scientific answers to a wide variety of questions regardless the origin and nature of these questions and the potentially huge distance between the questions and plausible answers. On the other hand, there are indeed intensified needs for analyzing and synthesizing what we know accumulatively and collectively about the nature of creativity and what we need to do in order to sustain and sharpen our competitive edge. Transformative research, for example, has attracted much of attention from an array of different types of stakeholders, notably, including the U.S. Congress, government and private funding agencies, uni-

[5] The search query used was: (influenza or flu) and (pandemic or epidemic or outbreak).

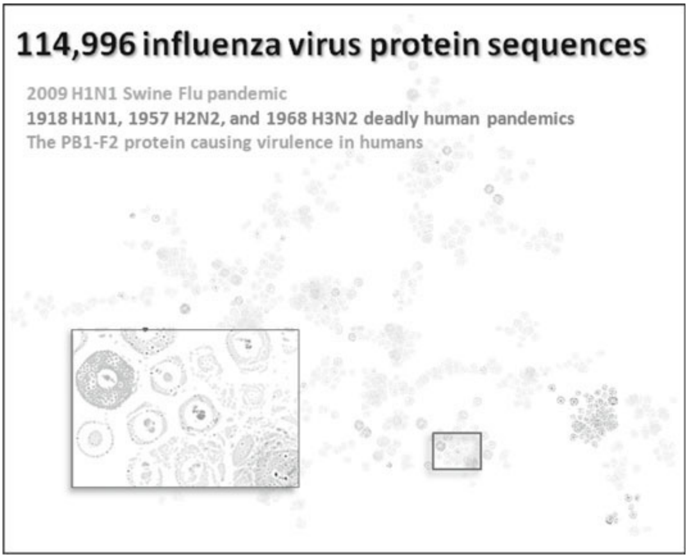

114,996 influenza virus protein sequences
2009 H1N1 Swine Flu pandemic
1918 H1N1, 1957 H2N2, and 1968 H3N2 deadly human pandemics
The PB1-F2 protein causing virulence in humans

Fig. 1.5 114,996 influenza virus protein sequences. Source: (Pellegrino & Chen, 2011) (see color figure at the end of this book)

versities, and individual scientists. What do we know about transformative research? How soon do we expect to recognize the transformative potential of a specific research plan? If transformative research is supposed to be so much ahead the state of the art, are existing assessment mechanisms, such as peer reviews and evaluations made by panels of established experts, capable enough of serving the role of a jury? What alternative and new mechanisms are there if the amount and complexity of information that we are supposed to examine goes beyond our reach? How do we handle false positives and increase the chance that truly transformative research gets recognized and supported?

Questions like these suggest that now we have more than enough reasons to study science — its history, present, and future — just about the same way science studies the nature and the mind. Calls for action made from policy makers and other stakeholders to science require an unprecedented level of accountability. There is a strong reason for developing evidence-based decision making and a new science of science policy. The notion of transformative research is at the spotlight of science policy and accountability as well as specific research planning for individual scientists. Supporting more transformative research is of critical importance in the fast-paced, science and technology-intensive world of the 21st Century.

There is an obvious lack of consensus on what exactly counts as transformative research. The National Science Foundation (NSF) in the U.S. defines transformative research in terms of a potential return of extraordinary outcomes, for example, revolutionizing entire disciplines, creating entirely new fields, or disrupting accepted theories and perspectives (NSF, 2007). The emphasis is clearly on the potential that may lead to revolutionary changes of disciplines and fields. In contrast, European perspectives tend to emphasize the role of high risks in the equation to justify the potential high impact. The term scientific breakthrough is often used by european researchers and officials when referring to transformative research.

The NSF has implemented several mechanisms to promote the funding of transformative research, or risky science. For example, the EArly-concept Grants for Exploratory Research (EAGER) funding mechanism aims to support exploratory work in its early stages on untested but potentially transformative research. The NSF also has a quick-response funding mechanism called the Grants for Rapid Response Research (RAPID) to deal with natural or anthropogenic disasters or other unanticipated events.

Fig. 1.6 shows a network of terms used in 63 NSF EAGER award abstracts in the IIS program between 2009 and 2010. A network like this can give a high-level picture of what is going on in a highly volatile environment. Terms, more precisely noun phrases, appear in these abstracts are grouped together based on how often they appear side by side, known as co-occurrences. Frequently co-occurred terms tend to form denser groups, whereas terms that rarely appear together tend to stay in separated groups. This is a commonly used

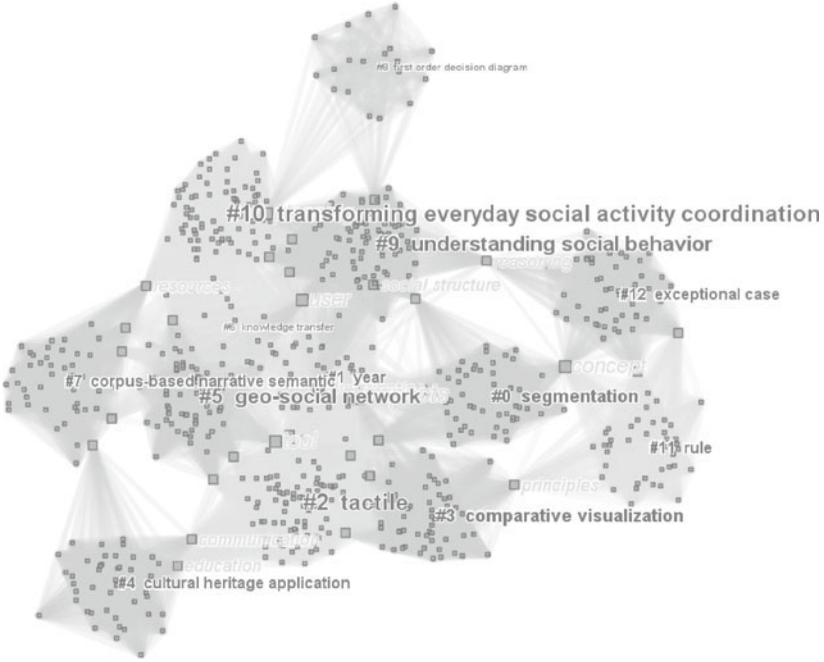

#8 first order decision diagram
#10 transforming everyday social activity coordination
#9 understanding social behavior
reasoning
resources
user
social structure
#12 exceptional case
#6 knowledge transfer
concept
#7 corpus-based narrative semantic
#1 year
#5 geo-social network
#0 segmentation
tool
#11 rule
principles
#2 tactile
#3 comparative visualization
communication
education
#4 cultural heritage application

Fig. 1.6 A network of 682 co-occurring terms generated from 63 NSF IIS EAGER projects awarded in 2009 (cyan) and 2010 (yellow). Q = 0.8565, Mean silhouette = 0.9397. Links = 22347. (see color figure at the end of this book)

technique to aggregate information so that we can identify emergent patterns at a higher level than the original information. In this book, we will use this type of thinking in our discussions. The aggregated groups are further labeled using broader terms so that we can make sense what each group is about. In this example, the labels of these groups — the terms in blue — are algorithmically chosen from the titles of these EAGER awards. It is often assumed that noun phrases are reasonable representatives of some underlying concepts. The occurrence of the term social behavior is interpreted as the evidence that the particular award involves the concept of social behavior. For example, transforming everyday social activity coordination is part of the title of an award that describes a lot of concepts of the group #10.

The assessment of the performance and accountability of the NSF in the U.S. is the responsibility of the Advisory Committee for GPRA Performance Assessment (AC/GPA[6]). The AC/GPA has about 20 members with substantial experience in academia, government, and industry. The 2009 AC/GPA membership list, for example, includes the Associate Vice President of the Office of Government and Community Affairs of University of Pennsylvania and the Dean of Rochester Institute of Technology's Golisano College of Computing and Information Sciences. The AC/GPA committee provides advice and recommendations to the NSF Director on its response to the reporting mandate required by the Government Performance and Results Act (GPRA) of 1993.

The AC/GPA evaluates outcomes from NSF's grant programs in research, education, and research infrastructure. The indicators take into account the support of potentially transformative research, stimulating innovation, developing successful models for teaching and learning, achieving active support of undergraduate and graduate students in research projects, and fostering research at large facilities or with advanced instrumentation that could not have been carried out without support from the NSF.

In addition to the AC/GPA, multiple mechanisms are in place for evaluating NSF's performance and accountability. Committees of Visitors (COVs) review program portfolios of NSF divisions every three years and provide external expert judgments regarding the quality and integrity of program operations and decisions and how research funded by the NSF have contributed to NSF's mission and strategic outcome goals. For example, the latest COV report of the Information & Intelligent Systems (IIS) Division in the CISE Directorate of the NSF was from May 19∼21, 2009[7]. The COV reviewed a total of 5,163 proposals, including 1,256 awarded and 3,907 declined proposals. This COV's members were from companies such as Google and Microsoft, and universities such as Stanford University, University of Toronto, and University of Washington.

The 2009 IIS COV paid special attention to how NSF IIS is able to support the current and prepare the next generation of innovators in the context

[6] http://www.nsf.gov/about/performance/acgpa/

[7] http://www.nsf.gov/od/oia/activities/cov/cise/2009/IIS%20COV%20Report.pdf

of the current global economic, social and climate conditions. The COV found that the division has a high quality and integrity of selecting and funding innovative and far-reaching research, although the amount of funding has not kept pace with the growing importance of IIS research. One of the questions addressed in the COV report was whether the program portfolio has an appropriate balance of innovative and potentially transformative projects. The COV identified several steps made by the IIS division in this direction:

- Specific instructions to the review panels to consider the transformative aspect of proposals
- Solicitations which push the frontiers of research
- Advice to panels to avoid implicit bias
- The creation of programs which require potentially transformative research

In response to the COV report, the IIS management acknowledged that low success rates continue to be a concern in each of the CISE divisions and in the NSF as a whole[8]. The NSF intends to make a strong case for increased investments in computing.

The Science of Science Policy and Innovation (SciSIP) Program at the NSF, directed by Julia Lane, is particularly relevant to issues concerning the performance evaluation of research typically at national and disciplinary levels. The growing portfolio of the SciSIP program includes a variety of innovative research projects that investigate technical and fundamental issues concerning evidence-based studies of the performance of science and research[9]. For example, one SciSIP project CREA[10] aims to develop measurements for analyzing highly creative research in the U.S. and Europe. The Cyber-enabled Discovery and Innovation (CDI) Program is an NSF-wide initiative on multidisciplinary research on innovations in computational thinking. As we shall see in this book, there are good reasons why interdisciplinary work may be an effective mechanism for transformative research.

1.5 Science and Society

A good understanding of a variety of views on the nature of the relationship between science and society is useful to set the context for many issues we will discuss in this book. For example, we will see where some of the most fundamental ideas of foresight seeking activities and the perspectives of a value-added chain model come from. We will be able to judge for ourselves whether we miss something significant and how things might look from an alternative view point. As a basic question about the nature of science, what

[8] http://www.nsf.gov/od/oia/activities/cov/cise/2009/IIS_Management_Response_to_the_COV_Report.pdf

[9] http://www.nsf.gov/about/performance/SciencePolicyWrkshp_Presentations/Tsuchitani.pdf

[10] http://www.cherry.gatech.edu/crea/

is the relationship between the producers and consumers of scientific knowledge?

The nature of the relationship between science and society is the subject of historiography. Two schools of thought have played significant roles in the evolution of our understanding of the subject: the internalist and externalist schools (Schuster, 2010). The internalist school, or internalism, is seen as focusing more on the cognitive aspects of science, whereas the externalist school, or externalism, is seen as focusing more on the socialeconomic dimensions of science. Internalists believe that science is autonomous in that the history of science is the development of pure thought over time and the development depends much more on the shoulders of geniuses than any contextual influence. In contrast, externalists believe that both the content and the direction of scientific knowledge were shaped by technological pulls that ultimately depended on economic and social needs. The extensive debate between internalism and externalism primarily focused on the nature of the cognitive inside of science and its social outside. For internalism, everything is intellectual — contexts were fine as long as they were seen as intellectual, not social. In contrast, for externalism, everything is social — the cognitive inside of science was fine as long as it was shaped by social factors.

One of the most influential internalists was Alexandre Koyré (1892–1964). To Koyré, the development of modern science depended on a revolutionary shift in ideas or theories. He produced a classic model of internalist explanation of the revolutionary origins of modern science and emphasized the critical role of a metaphysical framework as the only viable framework for scientific advance.

The early representative of externalism was generally regarded as Boris Hessen's paper delivered at the Second International Congress of the History of Science in 1931 in London. At that time, the work of Albert Einstein was under attack in the Soviet Union. Philosophers of the Soviet Union argued that Einstein's work was driven by bourgeois values and therefore it should be banned. Henssen's paper, one of the several papers delivered by the Soviet delegation, was entitled "*The Social and Economic Roots of Newton's Principia*." He asserted that Newton's work was inspired by his economic status and context, and that the Principia was little more than the solution of technical problems of the bourgeoisie. Hessen showed that scientific validity could exist regardless the origin of motivations. To Hessen, changes in the socioeconomic base produced the greatest achievement of the age, Newton's science.

Robert K. Merton, a historian and sociologist of science, substantially refined and further developed Hessen's work. Merton's work on the development of empirical and quantitative approaches that could demonstrate the influence of external factors on science has been seen as the precursor of scientometrics, the field of the quantitative study of science. To Merton, scientists' interests were ultimately driven by the internal history of the science in question. In other words, Merton's work showed early signs of how the contradictions between internalism and externalism can be resolved.

As Shapin (1992) pointed out, Merton's work went beyond refining and extending externalism — he was building a connection across over the cognitive/social barrier. He acknowledged that both internal and external factors played a role in the history of science.

To some, the internalist-externalist debate was resolved in the 1970s with the post-Kuhnian sociology of scientific knowledge and contextual history of science. A new interalism and a new externalism are mixed and evolved. To others, however, the debate is not over. The issue remains open.

1.6 Summary

The debates in the U.S. over the nature and extent of the crises and priorities of action have profound implications. They are among the most substantial proactive and responsive self-assessments since Pearl Harbor, Sputnik, and 911. These self-assessments are valuable and crucial for sustaining the competitive edge. The Yuasa Phenomenon and its potential causes are particularly interesting in this context. Emergent trends and patterns at macroscopic levels demand explanations at microscopic levels.

What is the role of creativity in scientific discovery and innovation?

What can be done to increase our creativity?

Regardless of our opinions in response to the specific arguments and interpretations of available evidence concerning the Gathering Storm, it is vital to sustain and enhance the competitive position of a country, the drive of a discipline, and the creativity of ourselves. Evidently, how to achieve such a goal is one of the top priorities on the agenda of a plethora of stakeholders from so many directions.

References

Augustine, N.R. (2007). Is America Falling Off the Flat Earth? : National Academies Press.

Benderly, B.L. (2010). Does the U.S. produce too many scientists? Scientific American, http://www.scientificamerican.com/article.cfm?id=does-the-us-produce-too-m&offset=6.

Collins, R. (1998). The Sociology of Philosophies: A Global Theory of Intellectual Change. Cambridge, MA: Harvard University Press.

COSEPUP. (1993). Science, technology, and the federal government: National goals for a new era. Washington, DC: National Academy Press.

Eysenck, H.J. (1993). Creativity and personality: Suggestions for a theory. psychological inquiry, 4(3), 147-178.

Lowell, B.L., & Salzman, H. (2007). Into the eye of the storm: Assessing the evidence on science and engineering education, quality, and workforce demand: The Urban Institute.

National Academy of Sciences, National Academy of Engineering, & Institute of Medicine of the National Academies. (2007). Rising above the gathering storm: Energizing and employing America for a brighter economic future. National Academies Press.

NSF. (2007, September 25). Important notice No. 130: Transformative Research. http://www.nsf.gov/pubs/2007/in130/in130.txt. Accessed 14 Aug 2010.

Pellegrino, D.A., & Chen, C. (2011). Data repository mapping for influenza protein sequence analysis. Proceedings of 2011 Visualization and Data Analysis (VDA). SPIE.

Schubert, A. (1987). Quantitative studies of science a current bibliography. Scientometrics, 12(5-6), 395-412.

Schuster, J.A. (2010). Internalist/Externalist Historiography, Encyclopedia of the Scientific Revolution from Copernicus to Newton.

Shapin, S. (1992). Discipline and bounding: the history and sociology of science as seen through the externalism-internalism debate. History of Science, 30, 333-369.

Yuasa, M. (1962). Center of scientific activity: its shift from the 16th to the 20th century. Japanese Studies in the History of Science, 1, 57-75.

Zhao, H., & Jiang, G. (1985). Shifting of world's scientific center and scientists' social ages. Scientometrics, 8(1-2), 59-80.

Chapter 2 Creative Thinking

What do we know about creativity? Where do insightful and enlightening moments come from? Are there such things as strategies and generic mechanisms for creative thinking and problem solving?

The general consensus of creative thinking is that we ought to think outside the box and that we should maintain an open mind as much as we can. However, a practical question is: What does it take to move from where we are now to the next — somewhat more desirable — position in the vast space of potential discoveries and solutions? We will revisit the navigation metaphor in Chapter 3. For now, let's focus on this question: are there any intriguing and tangible patterns that are generic enough from the diverse collection of the wisdom to get us started and help us move along?

2.1 Beyond Serendipity

We have all heard of stories of how a falling apple set Isaac Newton and his gravitational theory on the right track and how Henri Poincare arrived at his ultimate moment of enlightening just as he was stepping on a bus. Serendipity is one of the most fascinating, widely-admired, and yet most mysterious characterizations of creativity. Through the lens of serendipity, everything would magically fall into place so effortlessly that it leaves no trace and no clue of how one gets there. Discoveries rendered with the serendipitous paint make fascinating headline stories, and yet they improve little of our knowledge in terms of what we have to do to get there ourselves. Indeed, the notion of serendipity categorically denies rational pursuits of creative thinking.

The April issue of *PloS Biology* in 2004 reported a study of brain activity that accompanies the so-called 'Aha!' moments — the moments of inspiration.[1] Researchers gave participants a series of word problems to solve and studied their brain activities using brain imaging techniques. They found that activity increased in an area called the temporal lobe, in the right lobe

[1] http://men.webmd.com/news/20040413/scientists-explain-aha-moments

C. Chen, *Turning Points*

of the brain, when the participants reported experiencing creative insight. In contrast, little activity was shown in this area if no insightful experience was reported. Researchers have long suspected that the temporal lobe may play an important role in connecting distantly related information together. What is known about the type of creativity that connects disparate bodies of information?

A surprisingly rich variety of theories, models, and even tools exist in the literature, albeit sporadically laid out. Much of the available literature addresses the nature of creativity in terms of unique variations or focusing on specific phenomena in particular contexts. In contrast, one of the most persistent themes is the notion of making previously unknown connections. In the rest of the chapter, we review some of the most representative lines of research and pay special attention to the extent that the notion is embedded in a variety of seemingly unrelated theories and approaches.

Is it possible that a serendipitous arrival of insights is merely the final step of a subconscious process of searching for missing links? We often do not have a clear idea of what links are missing exactly until the idea becomes clear enough and that clear-enough idea forms the core of a discovery.

2.2 The Study of Creative Work

In a recent review of the study of creativity, Hennessey and Amabile (2010) found that the study of creativity is surprisingly fragmented, whereas research into the psychology of creativity has been rapidly expanding. Researchers in one subfield often seem unaware of advances in another. They call for more interdisciplinary research based on a systems view of creativity that recognizes a variety of interrelated forces operating at multiple levels.

Hennessey and Amabile sent survey questions to active researchers and theorists who have made the most significant contributions to the creativity literature. They asked 26 such researchers to nominate up to 10 “must-have” papers published since 2000 and heard back from 21 of them with over 110 nominated journal articles, books chapters, books, or special issues. To their surprise, there was “so very little” overlap between the 110 nominated works. Only seven of them were nominated by two people and only one was nominated by three people. The remaining nominations were plainly unique!

One of the first models of a creative process was proposed by Graham Wallas (1858 – 1932), an English social psychologist, in his 1926 work *The Art of Thought*. His model of creativity consists of five stages:

1) preparation
2) incubation
3) intimation[2]
4) illumination

[2] Some consider intimation as a component of a stage instead.

5) verification

In the first stage — preparation — the problem is identified and formulated. Previous work on the problem is also studied in this stage. The problem is then internalized in the incubation stage. There may be no apparent progress on solving the problem in this stage. Importantly, this period of interruption seems to be necessary for breaking away from misleading signals and false alarms. In the intimation stage, we can feel that a solution is on its way. In the illumination stage, the insight or the spark of creativity bursts through from its preconscious processing to conscious awareness. The insight often arrives suddenly and intuitively. Eventually, the idea is verified and evaluated. The question that many of us want to ask is: What does it take to be able to reach the illumination stage and find the inspirational insight?

Researchers and practitioners have repeatedly asked whether creativity is what we were born with or it can be trained and learned. The practical implications are clearly related to the fact that individuals in organizations are expected to become increasingly creative as they collaborate in project teams. A meta-analysis conducted by Scott and colleagues (2004) reviewed the results of 70 studies of creative training effects and found that carefully constructed creativity training programs typically improve performance. In contrast, Benedek and his colleagues (2006) studied whether repeated practice can enhance the creativity of adults in terms of the fluency and originality of idea generation. They found that while training did improve the fluency, no impact on originality was found.

The American psychologist Howard E. Gruber (1922 – 2005), a pioneer of the psychological study of creativity, questioned the validity of lab-based experimental studies of creativity. He argued that because creative works tend to be produced over a much longer period of time than the duration of a lab experiment, the laboratory setting is simply not realistic enough to study creativity. As a result, Gruber (1992) was convinced that an alternative approach, the evolving systems, should be used for the study of creativity. To him, a theory of creativity should explain the unique and unrepeatable aspects of creativity rather than the predictable and repeatable aspects seen in normal science.

Gruber strongly believed that the most meaningful study of creative work should focus on the greatest individuals rather than attempt to develop quantitative measures of creativity based on a larger group of people. His work, *Darwin on Man: A Psychological Study of Scientific Creativity*, is a classic exemplar of his evolving systems approach. His principle is similar to Albert Einstein's famous principle: as simple as it is, but not simpler. He strongly believed that characteristics of the truly creative work may not be found in an extended population (Gruber, 1992). Instead, he chose to study how exactly the greatest creative people such as Charles Darwin made their discoveries. He chose in-depth case studies over lab-based experimental studies.

Creativity is purposeful work. Gruber studied the lives of famous innovators and found broad common characteristics:

1) They engaged in a variety of activities within their chosen fields,
2) They held a strong sense of purpose about their work,
3) They had a profound emotional attachment to their work,
4) They tended to conceptualize problems in terms of all encompassing images.

The key to Gruber's approach is a radical focus on individuals as situated in a network of enterprise. His method uses a strong existential perspective as regards the "creative" individual who acts at all times with knowledge, purpose and affect. Lavery (1993) summarized Gruber's methodology as follows:

> "*It may well be the case that the seemingly random juxtaposition of ideas produces something new. But this juxtaposition arises in one person's mind. It is he who activates the structures giving rise to the ideas in question. It is he who recognizes the fruit of the encounter and assimilates it into a newly forming structure. And it was he in the first place who assembled all these constituents in the close proximity of one person's mind, his own, so that all this might happen*" ("And the Bush" 287)

So what factors may influence creativity? Numerous studies have looked into this question. In particular, a higher degree of variations in information is often associated with increased creativity. In fact, Maddux, Adam, and Galinsky (2010) found that recalling a multicultural learning experience obtained in abroad could improve the creativity of solving problems. The awareness of underlying connections and associations is also increased by such experiences. A different study found that using a single knowledge structure tends to increase the quantity of solutions, but using multiple knowledge structures tends to increase the quality and originality of problem solutions, especially combining either schema or associations with cases (Hunter, Bedell-Avers, Hunsicker, Mumford, & Ligon, 2008). On the other hand, in a marketing problem solving context, research indicates that early exposures to conflicting information may have a negative impact on creativity (Friedrich & Mumford, 2009).

The possible link between madness and creativity has been a topic of long interest in the literature of psychology, psychiatry and beyond, but the overall picture is still not clear. In 1998, a meta-analysis synthesized the results of 29 studies and 34 review articles on mental illness and creativity. The meta analysis found that although many authors asserted positive and causal connections, available scientific evidence was limited at the time (Waddell, 1998). The connection was reviewed again in a study published in 2004 (Lauronen, Veijola, Isohanni, Jones, Nieminen, & Isohanni, 2004). This time it was clear that evidence did exist, but the direction of any causal link was unclear. In 2006, Hungarian researchers further reviewed the scientific literature on the association of psychopathology and creativity. Contrary to the earlier focus on a strong association between schizophrenia and creativity, the current literature suggests that prominent social and artistic creativity is primarily associated with affective, and more specifically with bipolar affective illnesses. Being passionate is probably one of the necessary factors that would go hand in hand with creativity. As Nobel laureate Max Planck once said, "The creative scientist needs an artistic imagination."

2.3 Divergent Thinking

Contrary Imaginations was a citation classic written by the British psychologist Liam Hudson (1966). It was identified as a citation classic in the issue of *Current Contents* in October 1980. Hudson noticed that schoolboys seem to have different levels of abilities to handle *convergent* and *divergent* questions. A typical convergent question gives multiple possible answers for an individual to choose, for example:

Brick is to house as plank is to

a. orange
b. grass
c. egg
d. boat
e. ostrich

In contrast, a divergent question is an open-ended question that may have a numerous number of answers, like the question:

How many uses can you think of for a brick?

More interestingly, individuals differ considerably in terms of the number of answers they can come up with. Some can think of many different ways to use a brick, whereas others may be only able to think of one or two. Based on the ability of an individual to answer these types of questions, Hudson differentiated individuals in terms of two intellectual types: *convergers*, who would specialize in mathematics and physical sciences, and *divergers*, who are likely to excel in the arts and make surprising cognitive leaps.

The 1981 Nobel Prize in Medicine was awarded to Roger Sperry for his pioneering work on split-brain, which revealed the differences between hemispheres of the cerebral cortex and how the two hemispheres interact. The two hemispheres of the brain are connected. Each hemisphere has different functions. It is essential for the two hemispheres to communicate and function as a whole. If the connection between the two hemispheres is damaged, it results in a so-called split-brain. Studies of split-brain patients show that the left brain is responsible for analytic, logical, and abstract thinking such as speaking, writing and other verbal tasks, whereas the right brain is responsible for intuitive and holistic thinking such as spatial and nonverbal tasks. The right brain is also believed to home divergent thinking.

Divergent thinking has been widely regarded as a major hallmark of creativity. The distinction between convergent and divergent thinking was first made by the American psychologist Joy Paul Guilford (1897–1987), one of the pioneers in the psychometric study of creativity. He suggested that a prime component of creativity is divergent thinking (Guilford, 1967). The original terms were convergent and divergent production.

According to Guilford, divergent thinking is characterized by the presence of four types of cognitive ability (Guilford, 1967):

1) Fluency — the ability to produce a large number of ideas or solutions to a problem rapidly.
2) Flexibility — the ability to consider a variety of approaches to a problem simultaneously.
3) Originality — the tendency to produce ideas different from those of most other people.
4) Elaboration — the ability to think through the details of an idea and carry it out.

In contrast, convergent thinking is the ability to converge all possible alternatives to a single solution. When we take a test with multiple choice questions, we are typically using convergent thinking.

Following Guilford, divergent thinking is essential to creativity because it brings together ideas across different perspectives to reach a deeper understanding of a phenomenon. Divergent thinking can be easily recognized by its considerable variety and originality of the ideas generated. A large number of tests have been developed to measure the capacity for divergent thinking. Tests like *Alternate Uses* ask individuals to come up with as many different ways of using a common object as possible, such as a paper clip or a brick. Nevertheless, research has shown that one should be cautious in interpreting the implications of such measurements because this type of measurement may not adequately take the context of creativity into account. It is essential that creativity is studied in an appropriate context. After all, the ability of divergent thinking with a paper clip may tell us little about an individual's talent in music.

Divergent thinking is a central topic of creativity. Fig. 2.1 shows the position of the phrase *divergent thinking* in the relevant literature on creativity. The diagram shows a network of major terms extracted from the abstracts of 5,656 articles on creativity published between 1990 and 2010. These records were retrieved from the Web of Science in 2010 with a topic search of 'creativity.' The linkage represents the relative strength of how often two terms appear in the same abstract. The terms shown have the highest centrality scores, which measure the importance of their role in linking other terms together in the network. The position of the term divergent thinking is among the most essential ones in the network, which means that the importance of divergent thinking in creativity is widely recognized.

How should we deal with divergent thinking and convergent thinking in scientific discovery? Kuhn presented his view on this issue in a chapter of *The Essential Tension* (Kuhn, 1977), which was based on his 1959 speech at a conference on the identification of scientific talent. While recognizing the predominant attention to divergent thinking in scientific discovery, Kuhn argued that one should really take convergent thinking as well as divergent thinking

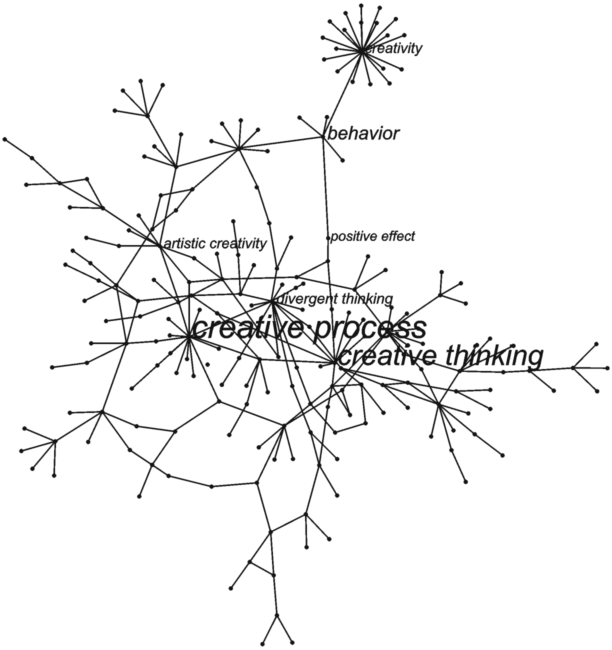


Fig. 2.1 A co-occurring network of major terms (noun phrases) extracted from the abstracts of 5,656 articles on creativity (1990 – 2010).

into account simultaneously because it is the dynamic interplay between the two forms of thinking that drives the scientific creativity. A common misconception of Kuhnian scientific revolutions is that they are rare and separated by prolonged uneventful normal science. The misconception still exists today despite the fact that Kuhn pointed out the misconception early on. In addition to ground-breaking and world-shaking scientific revolutions, scientists experience numerous conceptual revolutions at much smaller scales. Because divergent thinking plays a dominant role in initiating a change of world views and convergent thinking plays a crucial role in consolidating the new direction, the tension between the two forms of thinking is essential. Just as that divergent thinking is *yin* and convergent thinking is *yang*, the development of science is a holistic process of antithesis elements! This type of interplay can be seen in the following sections.

2.4 Blind Variation and Selective Retention

Many researchers have been deeply intrigued by the analogy between trial-and-error problem solving and natural selection in evolution. Is it merely an analogy on the surface or more than that? The American social scientist

Donald Campbell (1916–1996) was a pioneer of one of the most profound creative process models. He characterized creative thinking as a process of blind variation and selective retention (Campbell, 1960). His later work along this direction became known as a selectionist theory of human creativity. If divergent thinking is what it takes for blind variation, then convergent thinking certainly has a part to play for selective retention.

Campbell was deeply influenced by the work of Charles Darwin. In Campbell's *evolutionary epistemology*, a discover searches for candidate solutions with no prior knowledge of whether a particular candidate is the ultimate one to retain. Campbell specifically chose the word *blind*, instead of *random*, to emphasize the absent of foresight in the production of variations. He argued that the inductive gains in creative processes hinge on three necessary conditions:

1) There must be a mechanism for introducing variation.
2) There must be a consistent selection process.
3) There must be a mechanism for preserving and reproducing the selected variations.

Campbell's work is widely cited, including both supports and criticisms. As of July 2010, his original paper was cited 373 times in the Web of Science, and over 1,000 times on Google Scholar.

Dean Simonton's *Origins of Genius: Darwinian Perspectives on Creativity* (1999) is perhaps the most prominent extension of Campbell's work. His main thesis was that the Darwinian model might actually subsume all other theories of creativity as special cases of a larger evolutionary framework. Simonton pointed out that there are two forms of Darwinism. The primary form concerns biological evolution — the Darwinism that is most widely known. The secondary form of Darwinism provides a generic model that could be applied to all developmental or historical processes of blind variation and selective retention. Campbell's evolutionary epistemology belongs to the secondary form of Darwinism. Campbell's proponents argued that the cultural history of scientific knowledge is governed by the same principles that guide the natural history of biological adaptations. Simonton provided supportive evidence from three methodological domains: the experimental, the psychometric, and the historiometric domains.

Critics of Campbell's 1960 model mostly questioned the blind-variation aspect of the model. A common objection is that there would be too many possible variations to search if there is no effective way to narrow down and prioritize the search. The searcher would be overwhelmed by the enormous volume of potential variations. In contrast, the number of variations that would be worth retaining is extremely small. A scenario behind the *British Museum Algorithm* may illustrate the odds (Newell, Shaw, & Simon, 1958). Given enough time, what are the odds of a group of trained chimpanzees typing randomly and producing all of the books in the British Museum?

Campbell defended his approach with the following key points and argued that the disagreement between his approach and the creative thinking processes of Newell, Shaw, and Simon was minor matters of emphasis.

1) There is no guarantee of omniscience. Not all problems are solved. Not all excellent solutions archived.
2) One should not underestimate the tremendous number of thought trails that did not lead to anywhere. What has been published in the literature is a small proportion of the total effort of the entire intellectual community. What has been cited is an even smaller proportion of it.
3) Selective criteria imposed at every step greatly reduce the number of variations explored.

The course of problem solving may drift away from the original goal and lead to unexpected achievements such as solving a new problem. Campbell positioned his work as a perspective on creative thought processes rather than a theory of creative thinking. A perspective merely points to the problems, whereas a theory specifies underlying mechanisms and makes predictions. Nevertheless, one should ask whether creative thinking processes are predictable by their very nature. Recall Kuhn's dialectic view of the essential tension between divergent and convergent thinking, divergent thinking is crucial in the blind variation stage of the process, whereas convergent thinking is important in the selective retention stage.

Other critics of Campbell's work argue that his model is not broad enough to account for the full spectrum of creativity. Campbell's supporters defended that although the process may not be involved in all forms of human behavior and thought, Campbell had made a compelling case that all genuine forms of human creativity and invention are characterized by the process (Cziko, 1998).

Stephen Eick is a friend of mine. He is an entrepreneur and a researcher in the field of information visualization and visual analytics. He once told me a vivid business model. Every business idea is like a tomato. Throw the tomato at an imaginative wall, i.e. the market. Keep throwing more tomatoes to areas where tomatoes start to stack up. In this Tomato Model, the entrepreneur initiates blind variations, and variations are selected and retained based on the reaction of the market! In an analogy to Darwin's natural selection, a prolific scientist who publishes widely would become widely known than someone who is less prolific or publishes in few specialized journals. In terms of citations, a persistently productive author would have a better chance to get attention of the scientific community than a less prolific colleague.

Derek de Solla Price (1922 – 1983), the father of scientometrics, noticed that the most frequently cited articles tend to be the most recent ones in the citation network of scientific articles (Price, 1965). This immediacy implies that scientists have a relatively short attention span and they often pay more attention to recently published work than those published some years ago. Published ideas can be forgotten very quickly unless these ideas are picked up and carried on by someone to refresh the collective memory of the scientific community. Retention is a long-term process.

Price referred to a conjecture made by Burton and Kebler (1960) that the literature is made of two distinct types of publications with very different half-lives — the classic and the transient contributions. According to Price (p. 515), "the research front builds on recent work, and the network becomes very tight." He estimated about 30∼40 articles published before a citing article would constitute the research front relative to the citing article. We calculated the average number of references cited by a paper in several fields and found that the average number is 31 (see Table 2.1). It seems reasonable to expect a paper to cite the entirety of its research front in this sense.

Table 2.1 Average number of references cited per paper.

Topics	Records	Average Ref per paper	Max Ref per paper
Pulsars	1,048	13	200
Knowledge organization	4,444	14	331
Terrorism	1,732	21	168
String theory	7,983	38	182
Mass extinction	1,847	67	1,078
Mean		31	

Price arranged 200 articles on the topic of N-rays chronologically and used a matrix of citations (articles in a column, cite articles in a row) to depict the research front of the subject. It was clear that the research front consists of about 50 articles published prior to the citing article. Researchers are less likely to pay attention to papers published before these 50 or so papers. To cope with this immediacy effect and keep an idea constantly visible, scientists need to publish their work persistently. How long does it take for a paper to become obsolete? It depends on particular fields of study. The research front tends to move fast as a field begins to emerge. For instance, the initial discovery of pulsars was followed by fast-paced publications. Within the first 18 months of the discovery, the average half-life of papers was as short as weeks rather than months or years. Publish or perish!

2.5 Binding Free-Floating Elements of Knowledge

The blind variation and selective retention perspective is also evident in the work of the late Chinese scholar Hongzhou Zhao, although it is not clear whether his work was influenced by Campbell's work. Outside China, Zhao is probably better known for his work on the dynamics of the world center of scientific activities (Zhao & Jiang, 1985). With his education in physics, Zhao defined an element of knowledge as a scientific concept with a quantifiable value, for instance, the concepts of force and acceleration in Newton's $F = ma$, or the concept of energy in Einstein's $E = mc^2$. The mechanism for variation in scientific discovery is the creation of a meaningful binding between previously unconnected knowledge elements in a vast imaginary space.

The complexity of an equation can be quantified in terms of the “weight” of the elements involved.

The space of knowledge elements contains all the knowledge that has ever been created. The number of potentially valuable elements is huge. However, only a small fraction of all these bindings are potentially meaningful and desirable. For example, suppose there were a total of N elements of mechanics just before Newton discovered his second law. The number of candidate elements for binding would be the number of ways to choose the three elements F, m, and a. If there were Q ways to connect these elements with operations such as addition and subtraction, then Newton could expect to find his answer in $W = A_N^3 A_Q^1$ possible variations. In general, the number of paths to explore would be $W = A_N^e A_Q^{e-2}$, where e is the number of elements involved and Q is the number of operations. Furthermore, one can define the entropy of knowledge S as $ln(A_N^e A_Q^{e-2})$. Creative thinking is to reduce the entropy S by binding certain elements.

The entropy S has an alternative interpretation — it measures the potential of creation. The number of candidate elements in our knowledge varies with age and expertise. To acquire a new knowledge element, we need to have a thorough understanding of a concept. If an element is available for a possible binding with other elements, it is called a free floating element. A beginner would have a rather small number of such elements to work with, so the potential of their creativity is low. In contrast, established scientists would have a large amount of knowledge elements, but they are likely to be biased from their existing mental models and often fail to see new options for binding. They tend to interpret new things in terms of an old perspective or theory. As a result, the number of truly available free-floating knowledge elements is also quite limited for these established scientists. Zhao used this framework to argue for the existence of an optimal age when the highest entropy S is reached. At the optimal age, the scientist would have not only enough free-floating knowledge elements to work with but also the least amount of bias that could hinder their ability to see new ways of binding knowledge elements.

The above description of Zhao’s model identifies situations where the number of meaningful connections is overwhelmingly outnumbered by the size of a vast space of potentially viable candidates. Examples include the study of the origin of the universe in astronomy, searching for biologically and chemically constrained compounds in the vast chemical space in drug discovery, and searching for new connections between previously disparate bodies of knowledge in literature-based discovery. Campbell was right — searching for a small number of meaningful solutions in a boundlessly large space seems to be a common problem. Indeed, Zhao’s element-binding model is almost identical to Campbell’s theory of creative discoveries in science, i.e. a blind variation and selective retention process.

Zhao was searching for insights that could further explain what would make the center of scientific activities shift. For instance, how would the optimal age of a scientist influence the quality of element-binding activities? The center of scientific activity refers to the country that has more than 25% of noteworthy contributions of the entire world. The notion of the optimal age suggests that the majority of scientists at their optimal ages are most creative. From the knowledge element binding perspective, scientists who are younger than the optimal age may have the best memory, but they are probably inexperienced and have not seen enough examples. Therefore, their ability to see potential and alternative ways of binding elements of knowledge would not be as efficient as those who are in their optimal ages. Zhao found that when a country's social mean age and the optimal age distribution get close to each other, the country's science is likely on the rise. In contrast, when the social mean age drifts away from the optimal age distribution, the country's science is likely in decline. If a country is to become the center of scientific activity, its scientists must be in the best shape to bind knowledge elements in the most effective and creative way.

The strength of Zhao's knowledge element approach is that it can be used to identify a macroscopic movement of scientific activity worldwide. It focuses on the high-level description of a generic binding mechanism. It has practical implications for science policy of a country. Unfortunately, Zhao's approach has shortcomings shared by many statistical approaches. It does not lend itself as a concrete, step-by-step procedure that an individual scientist can pursue in daily activities. It does not give guidelines on how a specific binding should be conceived. The mechanism for making new variations is completely blind. Each variation does not make the next variation any easier. We have a problem!

As we shall see next, theories and models of creative thinking are more valuable if they can provide tangible principles and criteria for improving the quality and efficiency of our creative thinking as part of our decision making and problem solving activities. Janusian thinking, boundary spanning and brokerage mechanisms, the TRIZ model are among the most prominent representatives in the constructive category. They tell us how we might identify or create potentially fruitful routes towards achieving our goals. They provide concrete mechanisms of how to broaden our horizon!

2.6 Janusian Thinking

Janusian thinking is a special form of divergent thinking. Shakespeare's Hamlet was torn between two conflicting and mutually exclusive choices that he had to make: either *to be* OR *not to be*. In contrast, the idea of Janusian thinking is to find a new perspective that can satisfy both *to be* AND *not to be* simultaneously.

Janusian thinking is proposed by Albert Rothenberg (1996) in 1979 as a process that aims at actively conceiving multiple opposites or antitheses simultaneously. It is named after the Roman god Janus, who had two faces looking in opposite directions (see Fig. 2.2). As we shall see shortly Janusian thinking can be seen as a special type of divergent thinking and it can be used to generate original ideas systematically. In addition, an interesting connection between Janusian thinking and the work of the sociologist Murray S. Davis on why we find a new theory interesting is discussed in Chapter 4.

Fig. 2.2 Janus, the Roman god. Source: (The Delphian Society, 1913).

Rothenberg studied Nobel Prize winners, creative scientists and artists. He came to realize that this is the type of thinking used by Einstein, Bohr, Mozart, Picasso, and many others in their creative work. For instance, the notion of symmetry has served the role of a variation mechanism in many famous scientific and mathematical discoveries — such as Einstein's relativity and Bohr's complementarity. Rothenberg was convinced that most major scientific breakthroughs and artistic masterpieces are resulted from Janusian thinking.

Janusian thinking progresses through four phases over extended periods of time: 1) motivation to create, 2) deviation or separations, 3) simultaneous opposition or antithesis, and 4) construction of the theory, discovery, or experiment. Typically, Janusian thinking starts by asking: What is the opposite of a concept, an interpretation, or a phenomenon? In next phase, scientists start to break away from the work of other scientists. The deviation does not usually occur all at once but as an evolution from previous thought. In Phase 3, opposites are juxtaposed simultaneously in a conception and this conception is then transformed in Phase 4 and leads directly to the creative outcome. Phase 4 represents the construction of the full dimensions of the theory or discovery. Ultimately the creative scientists encapsulate an area of knowledge and relations when bringing and using these opposites together. The nature of the new conception is similar to a Gestalt switch or a change of viewpoint (see Fig. 2.3). Finding the right perspective is the key to creativity.

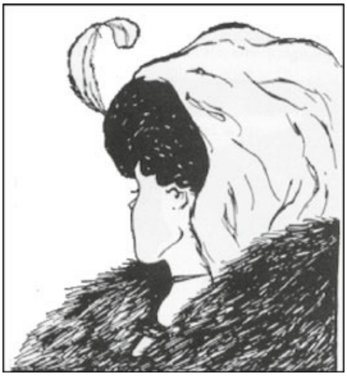


Fig. 2.3 The famous "my wife and my mother-in-law" by W. E. Hill (1915).

It is common to see that even after the critical conception is in place, there is usually still much of work to be done (Rothenberg, 1996). The construction phase includes components such as articulation, modifications, and transformations to produce integrated theories and discoveries. Articulation keeps elements separated but connected. In this sense, Janusian thinking is a brokerage or boundary spanning mechanism that connects opposites and antitheses. We will discuss the brokerage-based theory of discovery in more detail in the following chapters. The brokerage theory of discovery was originally proposed in (Chen, Chen, Horowitz, Hou, Liu, & Pellegrino, 2009).

Researchers have noticed that scientists who made great discoveries may experience solutions and discoveries before their ultimate discoveries. For Darwin, it all came together when he was reading Thomas Malthus's *Population* (1826). Malthus's main point was that an overpopulation of a species in a closed environment would eventually lead to a devastating destruction of the species due to competition for existence. Darwin had read Malthus's work several times and he can see how the very same struggle for existence increasingly enhanced the odds of survival of the species in its environment. Favorable and unfavorable variations of species have to be considered at once and formed a simultaneous antithesis. Darwin depicted each and every known species as separated and connected with one another through evolution. Malthus's work also inspired A. R. Wallace when he independently conceived an idea of natural selection. The ability to see how oppositions, favorable-unfavorable and increase-decrease, work together simultaneously turned out to be critical in both of the creative discoveries of the principle of natural selection.

The trail of Janusian thinking was also evident in Danish physicist Bohr's discovery of the complementarity principle. Bohr's thinking experienced a shift from the principle of correspondence to a breakthrough conception of simultaneous antithesis. The complementarity principle explained how light could be both wave and particle simultaneously. Two mutually exclusive appearances are necessary for a comprehensive description of the situation. Bohr discovered that the two concepts of wave and particle were a description of two complementary sides of a single phenomenon — light.

Rothenberg (1996) interviewed twenty-two Nobel Prize laureates in the fields of chemistry, physics, and medicine and physiology from Europe and the U.S. The interviews followed a systematic research protocol focused on in-progress creative work. *Einstein, Bohr, and Creative-Thinking in Science* (Rothenberg, 1987) included analyses of autobiographical accounts and work-in-progress manuscripts pertaining to the creative formulations and discoveries of outstanding scientists of the past, such as Bohr, Darwin, Dirac, Einstein, Planck, and Yukawa. Einstein's general theory of relativity was used as an example to show the nature of his first thought as an antithetical idea. As Einstein recalled that the idea was "*for an observer in free fall from the roof of a house, there exists, during his fall, no gravitational field ... in his immediate vicinity. If the observer releases any objects, they will remain, relative to him, in a state of rest.*" The question was how gravity can be seen as present and absent simultaneously.

All of the scientists interviewed by Rothenberg specifically spoke of the advantage of coming fresh into a new area and not being preoccupied by some of the biases and assumptions of that area. Breaking away from existing perspectives is critical in creative thinking. One of the commonly used mechanisms of creative imagination is to use distant analogies to escape the constraints of our existing mental models, especially at early stages of the creative process. Research has shown that scientists at the initial stage of a revolutionary discovery often use distant analogies, and as they develop a better grasp of the nature of the problem, these distant analogies are gradually replaced with analogies that are near to the problem at hand. In Chapter 4, we will discuss another relevant phenomenon called Proteus Phenomenon.

Rothenberg noted that creative people may not always realize that they are taking these steps in their thinking themselves, but the trails of these steps can be traced retrospectively. Rothenberg argued that Janusian thinking works because oppositions and antitheses represent polarities and extremes of a scale or category. Their explicit involvement in the discovery process provides a basis for storing and extending knowledge (Rothenberg, 1996). In other words, the extreme instances function just as landmarks for scientists finding their paths and impose constraints on potentially valuable variations. These landmarks are valuable because they clarify the nature and content of intermediary factors, and more importantly, the latent and previously unknown category.

From a philosophical point of view, Murray Davis proposed an intriguing framework to explain why we find some theories more interesting than others (Davis, 1971). In essence, a new theory would be regarded as interesting if it triggers a switch of our perspectives from one that we have taken for granted to a view that may contradict to what we believe. The caveat is that the new theory should not over do it. If it goes too far, it will lose our interest. The difference between Davis' framework and Janusian thinking is subtle but significant. In Davis' framework, when we are facing two opposite and contradictory views, we are supposed to choose one of them. In contrast, Janusian thinking is not about choosing one of the existing views and discarding the other. Instead, we must come up with a new and creative perspective so that it can accommodate and subsume all the contradictions. The contradictions at one level are no longer seen as a problem at the new level of thinking. It is in this type of conceptual and cognitive transformation that discoverers create a new theory that makes the co-existence of the antitheses meaningful.

The ability to view things from multiple perspectives and reconcile contradictions is in the center of dialectical thinking. The origin of dialectics is a dialog between two or more people with different views but wish to seek a resolution. Socratics, Hegel, and Marx are the most influential figures in the development of dialectical thinking.

According to Hegel, a dialectic process consists of three stages of thesis, antithesis, and synthesis. An antithesis contradicts and negates the thesis. The tension between the thesis and antithesis is resolved by synthesis. Each stage of the dialectic thinking process makes implicit contradictions in the preceding stage explicit. An important dialectical principle in Hegel's system is the transition from quantity to quality. In the commonly used expression, "the last straw that broke the camel's back", the one additional straw is a quantitative change, where a breakdown camel is a qualitative change. The negation of the negation is another important principle for Hegel. To Hegel, human history is a dialectical process.

Hegel was criticized by materialist or Maxist dialectics. In Karl Marx's own words, his dialectic method is the direct opposite of Hegel's. To Marx, the material world determines the mind. Marxists see contradiction as the source of development. In this view, class struggle is the contradiction that plays the central role in social and political life. In Chapter 1 we introduced how internalism and externalism differ in terms of their views of the nature of science and its role in the society. Dialectic thinking does seem to have a unique place in a diverse range of contexts.

Opposites, antitheses, and contradictions in Janusian thinking in particular and dialectic thinking in general are integral part of a broader system or a longer process. Contradictory components are not reconciled but remain in conflict; opposites are not combined, and oppositions are not resolved (Rothenberg, 1996). Opposites do not vanish; instead, one must transcend the tension between contradictory components to find a creative solution.

With regard to the 5-stage model of a creative process, the most creative and critical components of Janusian thinking are the transition from the third phase to the fourth phase, i.e. from simultaneous opposition to the construction of a new perspective. In Campbell's perspective of blind variation and selective retention, Janusian thinking proactively seeks antitheses as a mechanism for variation and imposes retention criteria for variations that can synthesize theses and antitheses.

2.7 TRIZ

TRIZ is a method for solving problems innovatively. The Russian acronym TRIZ is translated into English as the Theory of Inventive Problem Solving (TIPS). It was originally developed by the former Soviet engineer and researcher Genrich Altshuller (1926 – 1998). His earlier experience at the Naval Patent Office was believed to influence his development of TRIZ. The landmark work on TRIZ is *The Innovation Algorithm* (Altshuller, 1999). It was first published in 1969. Its English translation was published in 1999.

The Innovation Algorithm has three sections: technology of creativity, dialectics of invention, and man and algorithm. Altshuller analyzed different methods of technical creativity and he was convinced that people can be trained to be innovative. He developed 40 principles that can be used to resolve technical contradictions. To Altshuller, the main obstacles to creativity are psychological barriers. These obstacles can be overcome through a higher creative consciousness — a TRIZ mind.

The entire TRIZ method is built on the belief that the process of creativity can be learned (Altshuller, 1999). The process of creativity can be detected and made accessible to those who need to solve problems creatively. An "algorithm" or a recipe is available for invention. Altshuller described how an inventor follows a path of a trial-and-error search: "*Eventually, an idea emerges: 'What if we do it like this?' Then, theoretical and practical testing of the idea follows. Each unsuccessful idea is replaced with another, and so on.*" This is indeed, as you have recognized it, the same idea as Campbell's blind variation and selective retention.

Altshuller illustrated the challenges for an inventor in the diagram shown in Fig. 2.4. An inventor needs to reach the point of Solution from the point of Problem. The final location of the Solution point is unknown. Each arrow in the diagram represents a trial. None of the trials in the diagram so far leads to a viable path to a solution. The number of such failed trials can be very large, ranging from thousands and even tens of thousands without finding a satisfying solution.

The core insight from Altshuller is distilled as a set of 40 principles that one can follow in searching for paths to inventions. These principles form a systematic approach to the invention of new solutions and the refinement of existing solutions. It is designed to facilitate inventors to formulate problems in a way that is strikingly similar to Janusian thinking. In a nutshell, TRIZ is a constructive approach to creative thinking.

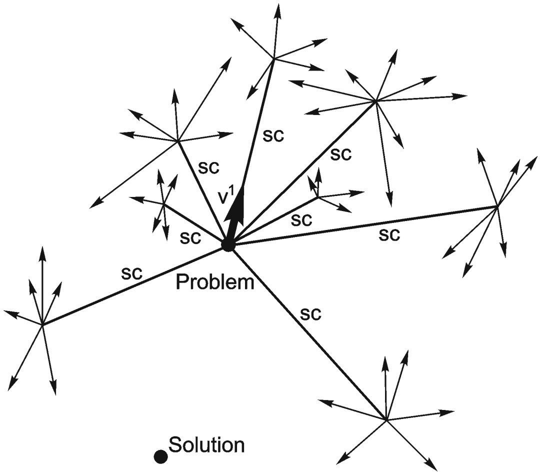


Fig. 2.4 Trial-and-error searches for a path to solution. Source: Figure 2 of (Altshuller, 1999), p. 25.

TRIZ focuses on technical contradictions and how to remove them. From this perspective, an invention is the removal of technical contradictions. Here are some examples of the 40 principles given in *The Innovation Algorithm*:

Segmentation:

1) Divide an object into independent parts.
2) Make an object sectional (for easy assembly or disassembly).
3) Increase the degree of an object's segmentation.

Extraction:

(Extracting, Retrieving, Removing)

1) Extract the "disturbing" part or property from an object.
2) Extract only the necessary part or property from an object.

Do It in Reverse:

1) Instead of the direct action dictated by a problem, implement an opposite action (i.e. cooling instead of heating).
2) Make the movable part of an object, or outside environment, stationary – and stationary part movable.
3) Turn an object upside-down.

Convert Harm to Benefit:

1) Utilize harmful factors — especially environmental — to obtain a positive effect.
2) Remove one harmful factor by combining it with another harmful factor.
3) Increase the degree of harmful action to such an extent that it ceases to be harmful.

The invention of light bulb illustrates how TRIZ works[3]. It was known

[3] http://www.salon.com/tech/feature/2000/06/29/altshuller/index.html

as early as 1801 that when electric current passes through metal filaments, filaments will light up. But the problem was that the filaments burned out in the process. The contradiction is that the filaments must get hot enough to glow but not too hot to burn themselves out. The contradiction was not resolved until 70 years later by the invention of Joseph Wilson Swan and Thomas Alva Edison. They solved the problem by placing the filaments in a vacuum bulb.

Another classic example is the design of tokamak for magnetic confinement fusion. Tokamaks were invented in the 1950s by Soviet physicists Igor Tamm and Andrei Sakharov. The problem was how to confine the hot fusion fuel plasma. The temperature of such plasmas is so high that containers made of any solid material would melt away. The contradiction was resolved by using a magnetic field to confine the plasma in the shape of a donut. Contradiction removal is a valuable strategy for creative thinking and problem solving in general. Focusing on a contradiction is likely to help us to ask the right questions.

2.8 Summary

Divergent thinking is widely recognized as one of the most essential characteristics of creativity. Therefore, how do we exactly achieve divergent thinking becomes a practical issue as well as theoretical one. To paraphrase Altschuller, it is not sufficient to know that we need divergent thinking and that we need to think outside the usual box.

Several theories and methods reviewed in this chapter provide a systematic strategy for finding solutions in a creative way. In contrast to the more commonly seen problem solving strategies such as divide and conquer, some of these creative thinking and problem solving strategies share a common principle — resolving contradictions. This means that we need to consider a contradiction or a pair of opposite problems simultaneously. Sometimes all it takes to turn a tough problem to a straightforward one is the change of our perspective!

Thomas Kuhn in his *The Essential Tension* argued that divergent thinking along is not adequate for the advance of science (Kuhn, 1977). He emphasized the role of convergent thinking in the stage of normal science. The nature of resolving a contradiction is synthesis. It is an uplifting of the current knowledge to a new level such that a contradiction to the old perspective is no longer a contradiction to the new perspective. Then it becomes clear that interplays between divergent and convergent thinking are integral part of creative thinking and problem solving.

The generic form of Darwinism echoes Kuhn's view on the essential tension between divergent and convergent thinking. What Janusian thinking and TRIZ have in common is their insight into the dynamics between theses and antitheses, and between contradictions and harmony. Dialectic thinking provides the most fundamental and consistent framework that subsumes many of the mechanisms for variation and creativity. It provides generic principles for keeping an open mind in the quest to new levels of the unknown.

References

Altshuller, G. (1999). Innovation algorithm: TRIZ, systematic innovation and technical creativity (1st ed.). Worcester, MA: Technical Innovation Center, Inc.

Benedek, M., Fink, A., & Neubauer, A.C. (2006). Enhancement of ideational fluency by means of computer-based training. Creativity Research Journal, 18, 317-328.

Burton, R.E., & Kebler, R.W. (1960). The 'half-life' of some scientific and technical literatures. American Documentation, 11, 18-22.

Campbell, D.T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. Psychological Review, 67(6), 380-400.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics, 3(3), 191-209.

Cziko, G.A. (1998). From blind to creative: In defense of Donald Campbell's selectionist theory of human creativity. Journal of Creative Behavior, 32(3), 192-209.

Davis, M.S. (1971). That's Interesting! Towards a Phenomenology of Sociology and a Sociology of Phenomenology. Philosophy of the Social Sciences, 1(2), 309-344

Friedrich, T.L., & Mumford, M.D. (2009). The Effects of Conflicting Information on Creative Thought: A Source of Performance Improvements or Decrements? Creativity Research Journal, 21(2-3), 265-281.

Gruber, H.E. (1992). The evolving systems approach to creative work. In D.B. Wallace & H.E. Gruber (Eds.), Creative People at Work: Twelve Cognitive Case Studies (pp. 3-24). Oxford, England: Oxford University Press.

Guilford, J.P. (1967). The Nature of Human Intelligence. New York: McGraw-Hill.

Hennessey, B.A., & Amabile, T.M. (2010). Creativity. Annual Review of Psychology, 61, 569-598.

Hill, W.E. (1915). My Wife and My Mother-in-Law. Puck, 16, 11.

Hudson, L. (1966). Contrary imaginations: A psychological study of the English schoolboy. New York: Schocken.

Hunter, S.T., Bedell-Avers, K.E., Hunsicker, C.M., Mumford, M.D., & Ligon, G.S. (2008). Applying multiple knowledge structures in creative thought: Effects on idea generation and problem-solving. Creativity Research Journal, 20(2), 137-154.

Kuhn, T.S. (1977). The Essential Tension: Selected Studies in Scientific Tradition and change. Chicago and London: University of Chicago Press.

Lauronen, E., Veijola, J., Isohanni, I., Jones, P.B., Nieminen, P., & Isohanni, M. (2004). Links between creativity and mental disorder. Psychiatry, 67(1), 81-98.

Lavery, D. (1993). Creative Work: On the Method of Howard Gruber. Journal of Humanistic Psychology, 33(2), 101-121.

Maddux, W.W., Adam, H., & Galinsky, A.D. (2010). When in Rome ... Learn why the Romans do what they do: how multicultural learning experiences facilitate creativity. Pers Soc Psychol Bull, 36(6), 731-741.

Newell, A., Shaw, J.C., & Simon, H.A. (1958). ELEMENTS OF A THEORY OF HUMAN PROBLEM-SOLVING. Psychological Review, 65(3), 151-166.

Price, D.D. (1965). Networks of scientific papers. Science, 149, 510-515.

Rothenberg, A. (1987). Einstein, Bohr, and creative-thinking in science. History of Science, 25(68), 147-166.

Rothenberg, A. (1996). The Janusian process in scientific creativity. Creativity Research Journal, 9(2-3), 207-231.

Scott, G.M., Leritz, L.E., & Mumford, M.D. (2004). The effectiveness of creative training: a quantitative review. Creativity Research Journal, 16, 361-388.

Simonton, D.K. (1999). Origins of Genius: Darwinian Perspectives on Creativity. New York: Oxford University Press.

The Delphian Society. (1913). The world's progress, Part III. Hammond: W. B. Conkey Company.

Waddell, C. (1998). Creativity and mental illness: is there a link? Can J Psychiatry. 43(2):166-172.

Zhao, H., & Jiang, G. (1985). Shifting of world's scientific center and scientists' social ages. Scientometrics, 8(1-2), 59-80.

Chapter 3 Cognitive Biases and Pitfalls

3.1 Finding Needles in a Haystack

Finding needles in a haystack is challenging because the haystack provides no clue at all about the needles. The Chinese culture describes such daunting situations with a similar metaphor but on a larger scale: finding a needle in an ocean!

Scientists, intelligence analysts, and many other decision makers in fact routinely need to deal with such situations. Astronomers who are searching for an earth-like planet in space, researchers working on drug discovery in pharmaceutical labs who are searching for the desirable compound in the vast chemical space, or intelligence analysts who need to connect the right dots among countless possibilities are just a few examples.

The difficulty of these types of problems can be measured by the ratio of the number of needles to the number of needle-like straws in the haystack. We can also formulate an index of difficulty in terms of the ratio of the number of earth-like planets to all the planets in the Universe, or the ratio of the number of desirable compounds to the total number of compounds in chemical space. In general, the problem solving literature uses the notion of an abstract space in which we search for paths to solutions to a problem. The ratio of the legitimate solutions to the size of the space, i.e. the number of candidates that we may have to consider, represents the magnitude of the challenge. In the worst case scenario, we may have to examine every inch of the entire space to find a satisfactory solution.

Scientists deal with the same type of space as they search for paths that will lead to scientific breakthroughs, new theories, and new understanding. However, the space of the unknown is more challenging because it may be an open space instead of a closed one and it may not even be coherent or consistent from one place to another. Solving problems and making new discoveries in such a space can be much more challenging than finding needles in a haystack. We cannot be certain if a needle is even in the haystack. We cannot be certain about when we can confidently stop the search without

C. Chen, *Turning Points*

knowing if we have found the best needle in the haystack yet. After all, we may not even have the foggiest idea of what the needle looks like. Scientific discoveries and creative thinking in general take place in such an open, dynamic space that is full of uncertainty.

3.1.1 Compounds in Chemical Space

Drug discovery is one of the characteristic examples of finding needles in a haystack. The needles here are compounds that have desirable properties for drugs, such as potency, and the haystack, a gigantic one, is chemical-biological space that pharmaceutical researchers need to work with.

It is generally accepted that modern drug discovery started with Paul Ehrilich's discovery of arsphenamine (Salvarsan). His discovery greatly improved the treatment of syphilis, a sexually transmitted disease. He made the discovery by systematically screening over 600 synthetic compounds. Today researchers routinely screen millions of compounds to find biologically active compounds.

Lipinski and Hopkins (2004), researchers of Pfizer Global R&D laboratories, intuitively described challenges in drug discovery using an analogy of navigating in the cosmological universe. The abstract space in which drug discovery navigates is a space of chemical compounds instead of stars and galaxies.

The chemical space is vast and diverse. Chemography is the technique for positioning compounds in the chemical space, just like the global positioning system (GPS) for locating streets and cities. Chemical biologists and drug discoverers are looking for effective ways to find those regions that are likely to contain biologically active compounds. The goal of the search is to find biologically relevant chemical space. An analogy in astronomy would be searching for planets in the universe that may have an atmospheric environment similar to the earth.

In the chemical space, the concept of distance between compounds is well defined, just as in the cosmological universe where the gravitational force keeps stars and galaxies in position. Similarly, compounds that are somewhat similar would be positioned nearer to each other than those less similar. There are many possible ways to measure the similarity between two compounds. This includes biological, physicochemical, and topological properties and even their proximity in the literature world — how often two compounds appear closely in text written by chemists, biologists or others. Each and every set of similarity measurements defines a metric space of compounds. Do these spaces have distinct structures? Are they compatible? Is it possible to map one space smoothly to the other? According to Lipinski and Hopkins, if compounds are positioned in the chemical space according to physicochemical properties, therapeutically useful compounds appear to cluster together in galaxies of compounds. However, some questions still remain: What are the implications of such tendency on discovering new drugs? How common is this tendency across the full spectrum of compound similarity measurement?

A far-reaching question is whether or not these galaxies of compounds are distributed evenly and sparsely — because an even and sparse distribution would make the search harder than otherwise. In the cosmological universe, it is known that the distribution is uneven. Much of the universe is empty, or void. If it is also true in the chemical space, galaxies of therapeutically interested compounds would be separated far apart by vast voids between them.

Thousands of high-throughput screening (HTS) programs suggested that compounds that bind to certain target classes are clustered together in discrete regions of chemical space. These regions can be defined by particular chemical descriptors.

Large pharmaceutical companies usually have files of 10^6 compounds. Chemical space is too large for a systematic scan. High-throughput screening serves as the starting point of the current primary strategy for the pharmaceutical industry for identifying biological active molecules.

HTS is one of the new concepts of drug discovery. A large number of hypothetical targets are simultaneously exposed to a large number of compounds. These compounds in turn represent numerous variations on a few chemical themes or fewer variations on a greater number of themes in HTS configurations. Hits in the HTS process are expected to become leads, the compounds that remain to be valid candidate in subsequent and more complex models. Data points are screening results associated with one compound at one concentration in a particular test. The number of data points has increased rapidly, from 200,000 at the beginning of the 1990s to the 5∼6 million in mid-1990s, and over 50-million around year 2000.

The leap-and-bounce increase has not generated any comparable increase in the research productivity of drug discovery. Although HTS has resulted in a large number of "hits," some industry leaders were disappointed that very few leads and development compounds, if any, can be credited to the new drug discovery paradigm (Drews, 2000). As pointed out by Jürgen Drews (2000): "The critical discourse between chemists and biologists and the quality of scientific reasoning are sometimes replaced by the magic of large numbers." Others have reached similar assessments (Lipinski & Hopkins, 2004), the generally poor quality of these data is not widely aware by those outside industry. Drug discoverers using HTS as a massive filtering mechanism need something else to improve the effectiveness of drug discovery.

Drug discovery is a lengthy process. It can take a decade or even longer from the initial basic research to its commercialization. Some discoveries are incremental, some are radical. Some discovered new core compounds for the first time, while others found new ways of using known compounds. The discovery process becomes increasingly expensive as it moves from initial research to clinical trials. A recently published study (Sternitzke, 2010) found that radical innovations are more likely than incremental ones to originate from basic research. On average, each drug is accompanied by 19 journal publications and 23 additional patents. Additional patent filings peak when the commercialization of the drug is in reach.

Kneller (2010) investigated the origins of 252 new drugs approved by the US Food and Drug Administration (FDA) between 1998 and 2007. He identified several factors that appear to play an important role in discovering innovative drugs, for example, the levels of public funding for academic biomedical research, rigorous peer review, and professional mobility. Higher levels of open, public funding are valuable for scientists trained in the course of academic research and biotechnology and pharmaceutical companies. Peer review in government funding agencies such as the NIH has been criticized for being reluctant to award funding to younger researchers and to non-traditional projects. However, the increased competitiveness required for successful proposals may nevertheless raise the quality of research and reduce the potential monopoly of senior professors as shown elsewhere, for example, in Japan. A higher professional mobility and career flexibility may contribute to more opportunities of cross-fertilizing ideas and initiatives.

As far as drug discovery is concerned, improving the efficiency of the process is a pressing problem. Navigators in the vast chemical space need signs and clues that can help them to choose new paths more efficiently. The current use of HTS is a truly blind variation mechanism. The mechanism thus far does not make use of any knowledge of the structural properties of the chemical space.

3.1.2 Change Blindness

We are facing the overwhelmingly profound challenges for finding paths in a variety of haystacks of evidence, oceans of data, and the universe of conjectures, hypotheses, mysteries and puzzles. In addition, our perception and cognitive system is vulnerable and prone to many biases and pitfalls. Some of the biases and pitfalls are so deeply rooted in our thinking, reasoning, and decision making that we often take them for granted without questioning their validity.

Our perceptual system is able to detect some patterns effortlessly. We can effortlessly sense the movement of an object. The ability to capture visual properties before we focus our attention is called preattentive perception. Preattentively achievable tasks include target detection, boundary detection, region tracking and counting and estimation. In general, we perform these tasks rapidly and accurately, in fact, in less than 200$\sim$250 milliseconds. It takes at least 200 milliseconds for us to make eye movements. Visual features such as color, shape, orientation, and motion are the best targets of preattentive perception. It is easy for us to notice a bright star in the dark sky, spot a red rose in a land covered by green vegetation, or pick "the odd one out" if it differs from the others in terms of its shape or size. However, our perceptual system is not very good at detecting and recognizing changes occurred in a scene, especially after our view is interrupted. The inability to detect this type of changes effectively is called *change blindness*.

One of the most well-known studies of change blindness was an experiment conducted by Simons (2000). In the experiment, the experimenter stopped a pedestrian on a street and asked for directions. The experimenter was holding a basketball as they were talking. Then a group of students walked by and interrupted their conversation and visual contact. During the interruption, the experimenter's basketball was taken away. After the brief interruption, the conversation was resumed. Very few pedestrians noticed that something was missing, but more than half of the pedestrians began to realize that the basketball was missing when they were asked specifically about it.

Change blindness can happen to both professionals and laymen. Researchers have come up with many theories in attempt to explain how and why change blindness happens. For instance, some theories focus on the role of stimulus and suggest that change blindness is due to the stimulus shown at different time. Some theories suggest that we only remember either what we see before or what we see after, but never both of them, so we have no way to compare and identify the difference between before and after. Theories arguing that we remember the before scene are known as the first impression theories, whereas theories that suggest we remember the after scene are known as overwriting explanation theories.

3.1.3 Missing the Obvious

In addition to change blindness, we tend to overestimate our perceptual and cognitive abilities. Colin Ware (2008) pointed out that one of the illusions we have is how much we are able to remember from a glance of a picture. We remember much fewer details from a picture than we realize. Although we can all agree that one picture is "worth a thousand of words," the fact is that much of what we see will vanish quickly from our perceptual memory and never get a chance to leave a trace in our memory. We cannot recall details that we did not get in the first place. Then why do we get the feeling that we see everything in the picture?

We can easily and in fact effortlessly direct our attention to any part of a picture, but at any time the area that actually gets our attention is rather limited. The reason that we feel like seeing all the details on the picture at a glance is because we could effortlessly attend to any spot on the picture if we want to. There is a significant difference between what we can readily achieve and what we have already achieved.

How reliable is an eyewitness's testimony? How reliable is our account of what we see in an unanticipated scene? The accuracy of eyewitness testimony is a topic of wide interest. Researchers studied the statements from 400~500 witnesses who were waiting for the arrival of President Kennedy in Dealey Plaza on November 23, 1963[1]. Researchers were interested in how accurately witnesses answer questions about what they saw or heard. How many gunmen were there in Dealey Plaza? How many gunshots were heard? Where were the gunshots originated? The study found that witnesses gave completely different answers to these questions.

The accuracy of eyewitness testimony is also questionable as details are revealed by recorded experiments. In one experiment, a staged assault of a professor took place in front of 141 unsuspecting witnesses. The incident was recorded on videotape. Immediately after the incident, everyone at the scene was asked to give detailed accounts. Most were inaccurate when compared to the videotape. Most overestimated the duration of the assault by two and half times longer. The attacker's weight was also overestimated, but his age was underestimated. After seven weeks, the witnesses and the professor himself were asked to identify the attacker from a group of photographs — 60% of them picked the wrong guy and 25% mistook an innocent bystander as the attacker.

The efficiency of our memory is characterized by the Yerkes-Dodson curve, which shows that our memory performs the best when we are neither too stressed nor too relaxed, but right in the middle of the two extremes. Many factors can influence our memory and even distort what we remember. Detectives are always keen on speaking to witnesses before witnesses have a chance to hear different versions of the story from other sources.

Researchers have been long interested in what separates the expert and from the novice in their ability to recall what they see and the role of expertise in explaining such differences. Chase and Simon (1973) conducted the best-known experiment concerning expert and novice chess players[2]. In their study, a chess master was regarded as an expert player. Between the chess master and a novice player, there was an intermediate-level player who was better than a novice player but not as good as a chess master. First, chess pieces were positioned according to a real middle game. All the players were given five seconds to view the positions. Then the chess board was covered. Players were instructed to reconstruct the positions of these pieces on a different chess board based on their memory. The performance was assessed in terms of the accuracy of their reconstruction. The master player performed the best, followed by the intermediate-level player. The master player correctly placed twice as many pieces as the intermediate-level player, who in turn correctly placed twice as many pieces as the novice player.

The second part of Chase and Simon's chess experiment was the same as the first one except that the positions of chess pieces were arranged randomly rather than according to real chess games. This time all players performed

[1] http://mcadams.posc.mu.edu/zaid.htm

[2] http://newton.nap.edu/html/howpeople1/ch2.html

with the same level of accuracy — the number of correctly recalled positions was much less than what the expert players could recall from settings based on real games.

Generally speaking, experts always demonstrate superior recall performance within their domain of expertise. As the chess experiments show, however, the superior performance of experts is unlikely to do with whether they have a better ability to memorize more details. Otherwise, they would perform about the same regardless whether the settings were random or realistic. So what could separate an expert from a novice?

Researchers have noticed that experts organize their knowledge in a much more complex structure than novices do. In particular, a key idea is chunking, which organizes knowledge or other types of information at different levels of abstraction. For example, the geographic knowledge of the world can be organized at levels of country, state, and city. At the country level, we do not need to address details regarding states and cities. At the state level, we do not need to address details regarding cities. In this way, we find it easier to handle the geographic knowledge as a whole. A study of memory skills in 1988 analyzed a waiter who could memorize up to 20 dinner orders at a time.[3] The waiter used mnemonic strategies to encode items from different categories such as meat temperature and salad dressing. For each order, he encoded an item from a category along with previous items from the same category. If an order included blue cheese dressing, he would encode the initial letter of the salad dressing along with the initial letters from previous orders. So orders of blue cheese, oil vinegar, and oil vinegar would become a BOO. The retrieval mechanism would be reflected at encoding and retrieval. Cicero (106 B.C. – 43 B.C.), one of the greatest orators in ancient Rome, was known for his good memory. He used a strategy called memory palaces to organize information for later retrieval. One may associate an item with a room in a real palace, but in general any structure would serve the purpose equally well. The chunking strategy is effectively the same as using a memory palace.

3.2 Mental Models and Biases

We can easily fall into some common pitfalls as we try to come up with new ideas or find unprecedented paths to discover the unknown. These pitfalls could hinder the quality of decisions we make or make us miss the target altogether.

We tend to pay more attention to things that immediately surround us than things that are further away. This tendency can be described in terms of the degree of interest. The intensity of our interest in a topic decreases as

[3] Ericsson, K. A. & Polson, P. G. (1988). An experimental analysis of the mechanisms of a memory skill. Journal of Experimental Psychology: Learning, Memory, and Cognition, 14, 305-316.

the distance between the point and topics that we are familiar with. This tendency may cause problems. While we tend to search locally, the real solution to a problem may take an extensive search to find.

Another type of pitfall is that we tend to take paths of the least resistance rather than paths that are likely to lead to the best answers or the best decisions. The more we go down the same path, the less resistance the path appears to be. Whenever a new path competes with a well-trodden one, our decision tends to be biased towards the familiar and proven path. We prefer no uncertainty and want to avoid unforeseen risks. This preference may have serious consequences!

This problem can be better explained in terms of mental models, or cognitive models. Mental models are simplified abstractions of how a phenomenon, the reality, or the world works. We use such models to describe, explain and predict how things work and how situations may evolve. However, we are also biased by our own mental models.

Mental models are easy to form and yet hard to change. Once we have established our mental model, we tend to see the world through the same mental model and reinforce it with new evidence. If we have to deal with evidence that apparently contradicts the model, our instinct is to find an interpretation for the contradiction rather than to question the model itself. Once we find extenuating reasons to convince ourselves why it makes sense that the evidence doesn't appear to fit the model, we move on with an uncompromised faith in the original model. Fig. 3.1 shows a series of drawings that gradually change from a man's face to a sitting woman. If we start to look at these images one by one from the top left, we see images showing a man's face until the last few images. In contrast, if we start it from the lower right and move backwards, we see more images showing a woman sitting there.

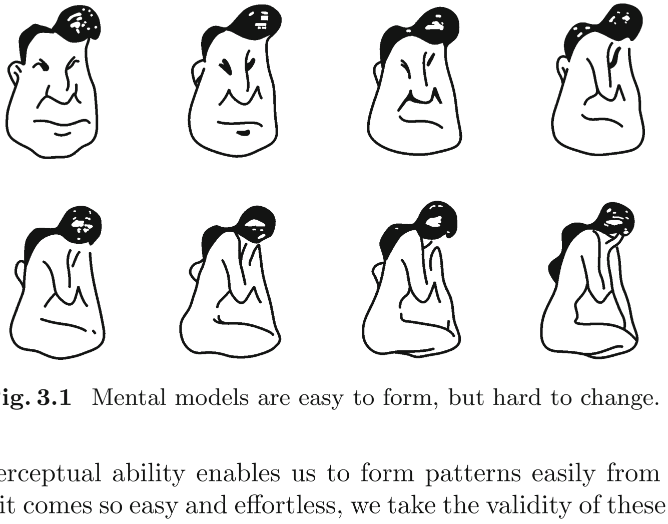


Fig. 3.1 Mental models are easy to form, but hard to change.

Our perceptual ability enables us to form patterns easily from what we see. Since it comes so easy and effortless, we take the validity of these patterns for granted. Sometimes such patterns prematurely narrow down the solution space for subsequent search. Sometimes one may unconsciously rule out the ultimate solutions. A simple connecting-the-dot game illustrates this point.

In this game, 9 dots are arranged in three rows and three columns (see Fig. 3.2). You are asked to find a way to connect all these dots by four straight lines. The end point of each line must be the starting point of the next line.

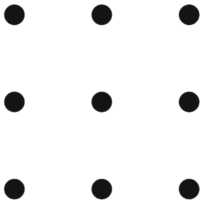


Fig. 3.2 Connecting the dot with no more than four jointed straight lines.

If there is still no sign of a solution after a few trials, ask yourself whether you are making any implicit assumptions and whether you are imposing unnecessary constraints to your solutions. The problem is usually caused by implicit assumptions we make based on a Gestal pattern that we may not even realize it is formed. These implicit assumptions set the scope of our subsequent search. In this case, we won't be able to solve the problem unless the implicit assumptions are removed. This type of blind spot is more common in our thinking than we realize. Sometimes such blind spots are the direct source of accidents.

Charles Perrow published a book called *Normal Accidents: Living with High-Risk Technologies* (1984). He made a series of compelling cases to underline the core insight that many accidents are caused by human factors. One of the cases was about an accident in the Chesapeake Bay in 1978. The captain of a Coast Guard cutter training vessel saw a ship ahead. It was dark and he saw two lights on the other ship, so he thought that was a ship going in the same direction as his own ship. What he didn't know was that there were actually three lights and it was going towards them. Since he missed one of the lights, his understanding of the situation was wrong. As both ships traveled at full speed and they were closing up rapidly, the caption misinterpreted that it must be a very slow fishing boat and he was about to overtake the boat. The other ship was, in fact, a large cargo ship. Both of them approached the Potomac River. The Coast Guard captain suddenly thought, still based on his incorrect mental model, he had to make a left turn so that the small and slow fishing boat could turn to the port. Unfortunately, this turn put the ship on a collision course with the oncoming freighter. Eleven coast guards on the ship were killed in the accident.

The captain's mental model in this case was how he perceived, interpreted and responded to the situation. He started with a wrong mental model and made a series of wrong decisions without questioning the model itself. At a larger scale, we form mental models of the reality not only individually, but also collectively. A mental model of the reality can be shared by a large number of people. A group of scientists can share a common scientific theory. A group of thinkers can share a school of thought. In Kuhn's *Structure of Scientific Revolutions*, a paradigm is accepted by a scientific community. It functions just like the mental model of the scientists in the community. For scientists who work in a well-established paradigm, it may become increasingly difficult to adopt an alternative paradigm and see how the world can be interpreted in a different way. Kuhn used the notion of Gestalt switch to explain the difficulty in changing perspectives.

How can we tell which models or paradigms are superior? Can we expect that paradigms getting 'better' all the time? Even if mental model is shared by a large number of intelligent people, it could still be a poor representation of the reality. The Ptolemaic system of the solar system is a classic example[4]. The Copernican model is not only a more accurate representation of the reality, but also a much simpler one.[5] The simplicity stands out when you look at the two models side by side in Fig. 3.3. Simplicity is one of the few criteria that we expect to see in a superior theory.

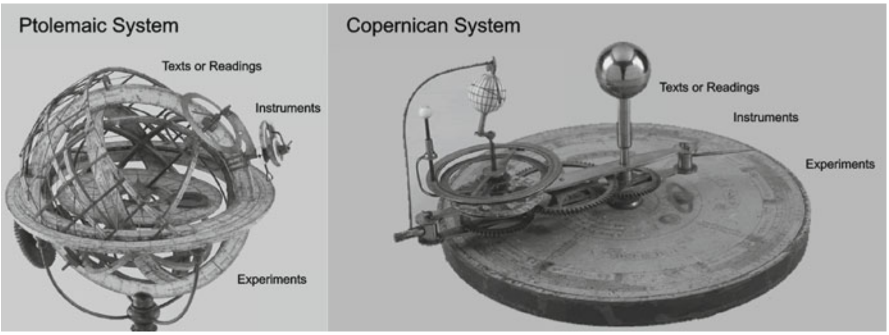


Fig. 3.3 Models of the Ptolemaic system (left) and the Copernican system (right).

Mental models and theories in general are about how the reality works or how a phenomenon takes place. They can be used to make predictions. Such predictions have a direct influence on what decisions we make and which course of action we take. There are two broad categories for us to assess the quality of a mental model or a theory. We can examine the coherence and integrity of a theory internally. A theory is expected to explain the mechanisms of a phenomenon in a consistent manner. We can also examine the utility of a theory externally. Does it complete with an alternative theory? Can it give simpler explanations of the same phenomenon than its competitors?

Don Norman, a pioneer of human-computer interaction, proposed that one should check the reality through seven critical stages of an action cycle. Norman's action cycle starts with the intended goal to achieve, proceeds to

[4] http://microcosmos.uchicago.edu/microcosmos_new/ptolemy.html
[5] http://microcosmos.uchicago.edu/microcosmos_new/copernicus.html

the execution of a sequence of acts to reach the goal, and an evaluation of the effect of the action with respect to the original goal. Norman particularly identifies the execution and evaluation as the two most critical stages. Norman's suggestion was made primarily for human-computer interaction, especially in situations where an end user of a device or a system needs to figure out how the system is supposed to react to the user's action without knowing the exact design of the system. Based on the appearance and layout of various controls of the system and the results of initial trials of a few controls, the user may develop an idea how the system works. This type of understanding becomes the user's mental model. The focus of a mental model can be a situation as well as a device. 911 terrorist attacks and the mass destructive weapon are examples of incorrect mental models of situations at much larger scales (Betts, 2007).

The feedback from the system is critical for the user or the analyst to assess their mental model. For some systems, the analyst has to probe the systems to find out how the system would react to a particular input. The time elapsed for the user to receive feedback from the system is also critical, especially for interactive systems. If it takes too long for the user to receive feedback, it is less likely that the user would be able to make use of the feedback and refine the mental model. For many real-world systems, however, it is impossible to get any immediate feedback, for example, the long-term effects of genetically modified food on human beings and human activities on climate change.

Our mental models may cause a different kind of change blindness. A stabilized mental model is also known as a mindset. The stability comes with both pros and cons. The advantage is that a stable mental model provides us a familiar framework to solve routine problems or to apply the same methodology repeatedly. As we become more familiar with the framework, we become an expert and specialized in this particular area of knowledge and our performance becomes more efficient. However, a mindset is resistant to change. We often take our own mental model for granted. It becomes harder and harder for us to take a fresh perspective and question the validity of our own mental model than revise and modify our mental models. The most serious drawback of a mindset is that we become increasingly biased and less open-minded. We tend to see everything through the same perspective and we hardly question whether our mental model is appropriate in the first place.

Specialists or established experts could be blinded by their own mindsets and fail to recognize something obvious from outsiders' perspectives. Because they are so knowledgeable of what might be expected, they may not be able to recognize unexpected patterns. As Kuhn pointed out, scientists who introduced new paradigms tend to be either an experienced scientist entering a new field or a young researcher at the early stage of his/her career. Scientists who resisted a new paradigm tend to be the most established leading experts. The reunification of East and West Germany surprised many specialists in the Central Intelligence Agency (CIA) (Heuer, 1999). Our perceptive and cognitive system can perform many complex tasks efficiently. However, we should bear in mind that it is also vulnerable in many ways that may influence the quality of our decision making and problem solving.

Heuer (1999) identified the inherent limitations on analysts' mental machinery. He urged the CIA that attention must be paid to techniques and tools for coping with these inherent limitations of human analysts. He recommended several steps that the CIA should take: promote and reward critical thinking, stay abreast of studies on how the mind works, and foster the development of tools to assist analysts in assessing information.

3.2.1 Connecting the Right Dots

Why do we often jump to conclusions? One possible reason is that we tend to underestimate the complexity of the nature or the world in general. If two events take place one after another, we tend to assume that the first event somehow causes the second one. When we need to find an explanation of what is going on, we often settle with the first good enough reason that we can find. In reality, in both types of situations, we may just pick the wrong end of the stick, so to speak.

As an urban legend says,[6] a family complained to General Motors' Pontiac Division about a strangely behaved engine of their new Pontiac. The family had a tradition of having ice cream for dessert after dinner every night. The whole family would vote on which flavor of ice cream and the father would drive his new Pontiac to the store to get it. Strangely enough, trips for vanilla ice cream always ran into the same problem — the car won't start on the way back, but if he gets any other kind of ice cream, the car would start just fine.

Pontiac sent an engineer to check it out. The engineer rode the car with a few ice cream trips. The car stalled on vanilla ice cream trips, and started on other flavors, just the way it was complained. Was the car allergic to vanilla ice cream?

The engineer noted a variety of details of the trips such as the type of gas used and the time to drive back. Then he had a clue: It took less time to buy vanilla than any other flavor. Because vanilla ice cream is popular, the store puts it at the front of the store and it is easy to get. It takes much longer to get other flavored ice cream because it is located in the back of the store. Now the question for the engineer was why the car wouldn't start when it took less time. The engineer was able to locate the problem. It was with the vapor lock. The engine needs enough time to cool down.

Richard Betts is a former member of the Military Advisory Panel of the Director of the Central Intelligence and of the National Commission on Terrorism. He compared two cases of intelligence failure — 911 terrorist attacks

[6] http://www.snopes.com/autos/techno/icecream.asp

and Iraq's missing weapon of mass destruction (WMD): in one case, the intelligence community failed to provide enough warning; in the other, it failed by providing too much (Betts, 2007).

It was commonly believed after the 911 terrorist attacks that U.S. intelligence had failed badly. However, Betts pointed out the issue is not that simple. The intelligence system did detect that a major attack was imminent weeks before the 911 attacks. The system warned clearly that an attack was coming, but could not say where, how, or exactly when. The vital component lacked from the warning was the actionability — it was too vague to act upon.

According to Betts, two months before 911, an official briefing warned that Bin Laden "will launch a significant terrorist attack against the U.S. and/or Israeli interest in the coming weeks. The attack will be spectacular and designed to inflict mass casualties." George Tenet, the Director of Central Intelligence (DCI), later said in his memoirs that "we will generally not have specific time and place warning of terrorist attacks." In addition, many intercepted messages or cryptic warnings were not followed by any terrorist attack. Before 911, more than 30 messages had been intercepted and there was no terrorist attack. Furthermore, it is not unusual to choose not to act on warnings if various uncertainties are involved. An extreme hurricane in New Orleans had been identified by the Federal Emergency Management Agency (FEMA) long before the Hurricane Katrina arrived in 2005. Fixing New Orleans's vulnerability would have cost an estimated $14 billion. The perceived cost-effectiveness before the disaster was not in favor of making such investments while its benefit was hard to estimate. The question sometime is not how to act upon a credible assessment of a threat or a potential disaster; rather, it is prudent to ask whether one should act given the cost-effectiveness in the situation. Gambling sometimes pays off. Other times it will be seen as a failure, especially on the hindsight. In Chapter 4, we will discuss the gambling nature of almost all decision-making in terms of the optimal foraging framework.

Another factor that contributed to the failure was due to the lost of focus caused by the tendency of maximizing collection. The trade-off between collecting more dots and connecting the dots is the issue. The fear of missing any potentially useful dots and the fear of taking direct responsibilities were driving the maximum collection of dots. However, collecting more dots makes connecting the dots even harder. Indeed, after reading the 911 Commission's report, Richard Posner concluded that it is almost impossible to take effective action to prevent something that has never occurred previously. In many ways, this is a question also faced by scientists, who are searching for ways to find meaningful dots and connect them. As we will see in later chapters, there are pitfalls as well as effective strategies for dealing with such situations.

The 911 Commission recommends that the dots should be connected more creatively and it is "crucial to find a way of routinizing, even bureaucratizing, the exercise of imagination." Mechanisms for thinking outside the box should be promoted and rewarded, just as Heurer (1999) recommended.

If American intelligence failed to connect the dots before 911, it made the opposite mistake on Iraq by connecting the dots too well. Betts showed that ironically policymakers paid little attention to intelligence on cultural and political issues where the analysis was right, but they acted on technical intelligence about massive destruction weapons, which was wrong.

The consensus after the first Gulf War was that Iraq had attempted to develop WMD. The lack of evidence that they had indeed destroyed the necessary facilities was interpreted as a sign that they were hiding from the inspections of the West. The mindset of the intelligence and policymakers was that the Iraqis concealed WMD as they did with their chemical and biological weapons prior to the first Gulf War. The lack of direct evidence of the existence only reinforced the model rather than questioned the assumption. After all, the lack of evidence was consistent with the hypothesis that Saddam was hiding it from the world. In addition, negative evidence that pointed to different scenarios did not get enough attention. It is known from psychology that people tend to maintain their existing mental models rather than challenge them when facing new evidence. It takes much stronger evidence, i.e. wake-up calls, to make people alter their mental model. In this case, analysts did not ask whether Iraq had WMD; instead, they wanted evidence that Iraq did not have WMD.

Given this mindset, conclusions were drawn from Iraq's behavior in the past rather than from any direct evidence of the existence from the present. For instance, documents obtained by the United Nation inspectors showed that an Iraqi government committee once gave instructions to conceal WMD activities from inspectors. At the end of the first Gulf War, Iraq admitted having chemical and biological weapons and claimed that they had destroyed them later, but never produced any evidence of such destructions. Misinterpretations of all available evidence and the lack of evidence only became obvious on the hindsight. Dots were connected without beyond-the-doubt evidence. True connections were slipped through the analysis.

On the hindsight, Betts suggested what the intelligence should have done given what was known and what was possible to know at the time. First, Iraq was probably hiding WMD weapons. Second, the sources that led to the conclusion were deductions from the past history. And third, there was very little direct evidence to back up the deduction. One of the lessons learned from the two opposite cases is a caution that one should not draw too many lessons from a single failure.

Pearl Harbor: Warning and Decisions by Roberta Wohlstetter (1962) was regarded as the first intelligence analysis that differed significantly from prior studies of surprise attacks. Wohlstetter's focus was on Pearl Harbor, but his insight is far-reaching. He studied Pearl Harbor in a much broader context than previous studies of similar surprise attacks would do. The key point he made was that analysts and decision makers in crisis situation like Pearl Harbor can be seriously biased by our own perception, or rather, misperception. The Pearl Harbor surprise was not due to a lack of relevant intelligence data.

Instead, it was due to misperceptions of the available information.

Abraham Ben-Zvi (1976) extended this line of analysis of surprise attacks and paid particular attention to both strategic and tactical dimensions. He analyzed five cases of intelligence failure to foresee a surprise attack and found that whenever strategic assumptions and tactical indicators of attack converged, an immediate threat was perceived and appropriate precautions were taken. When discrepancies emerged, people chose to rely on strategic assumptions and discard tactical indicators as noises.

As Heuer (1999) pointed out, it is essential to be able to recognize when our mental model needs to change. The weaknesses of human perception and cognition have been identified in numerous cases across a diverse range of situations. Nevertheless, we still need to promote the awareness of these weaknesses and remind ourselves that conflicting information may be an early sign for re-assessing what we think we know about a situation. The 'noises' may contain vital clues.

3.2.2 Rejecting Nobel Prize Worthy Works

The Nobel Prize is widely regarded as the highest honor and recognition of one's outstanding achievements in physics, chemistry, medicine, literature, and peace. In his will, Alfred Nobel described the prizes should be given to persons who have made the most important discoveries, the most outstanding work, or have done the best work for promoting peace regardless their nationality:

> "*one part to the person who shall have made the most important discovery or invention within the field of physics; one part to the person who shall have made the most important chemical discovery or improvement; one part to the person who shall have made the most important discovery within the domain of physiology or medicine; one part to the person who shall have produced in the field of literature the most outstanding work in an ideal direction; and one part to the person who shall have done the most or the best work for fraternity between nations, for the abolition or reduction of standing armies and for the holding and promotion of peace congresses.*"[7]

Although there is no question that the Prizes should be given to the most important discoveries and the most outstanding work, there are always discrepancies, to say the least, on which discovery is the most important and which work is the most outstanding. While the importance of some of the Nobel Prize winning work was recognized all along, the significance of other Nobel Prize winning work was overlooked or misperceived. How often

[7] http://nobelprize.org/alfred_nobel/will/will-full.html

do experts recognize the potential of important work? How often do they miss it? Why and how?

Peer review is a long established tradition in science. Scientists publish papers. Scientists seek funding to support their work and submit grant proposals to funding authorities. Peer review plays the most critical role in both scientific publication and funding allocation. There is a consensus, at least in the scientific community, that peer reviewed publications or grant proposals have a more prestigious status and a better quality than non-peer reviewed ones. In the peer review system, scientists submit their work for publication. The decision on whether their work should be published heavily relies on the outcome of reviews produced by peer scientists who usually work in the same field of study. Reviewers usually make recommendations whether a manuscript is publishable and whether additional changes are necessary before it is accepted for publication. Reviewers and authors may or may not be anonymous to each other. The anonymity can be in one way only or in two ways, which is known as either a single blind or a double blind process. Double blind peer reviews are regarded as more rigorous than other forms of peer reviews. Peer review of grant proposals works similarly.

As one can imagine, peer reviews have made blunders in both ways: blocking the good ones and letting the bad ones through. Tim Berners-Lee's paper on the idea that later led to the World Wide Web was rejected by the ACM Hypertext conference on the grounds that it was too simple by the standard of the research community. Reviewers of grant proposals were criticized to be too conservative to judge transformative research and high-risk and high-payoff proposals.

Rejecting a future Nobel Prize winning discovery would be a big enough blunder to get everyone concerned. If an achievement can be ultimately recognized by a Nobel Prize, how could experts misjudge its potential in its early days?

One of the most commonly given reasons for rejection in general is that the work in question is premature. Spanish scholar Juan Miguel Campanario noted a lack of interest from sociologists, philosophers and historians of science on how and why some important scientific discoveries were rejected, resisted, or simply ignored by peer scientists. Keeping an open mind is one of the most fundamental values upheld by a scientist. However, reviewers must also consider risks and uncertainties. In order to better understand what transformative research is and how one can recognize it timely, it is important to understand not only how scientific breakthroughs get instant recognition but also how revolutionary discoveries can be rejected, resisted, and ignored. Campanario identified some common patterns of resistance to scientific discovery:

- Papers are rejected
- Discoveries are ignored by peer scientists
- Published papers are not cited
- Commentaries oppose new discoveries

Campanario analyzed commentaries written by authors of highly cited papers and found that some of the most cited papers experienced initial rejections. It is an interesting phenomenon that some of the rejected papers became highly cited later on. Campanario then studied 36 cases of rejected Nobel class articles and 27 cases of resisted Nobel class discoveries and concluded that there is a real danger and the consequence can be disastrous.[8] He searched through autobiographies of Nobel Prize winners and conducted a survey among Nobel Laureates awarded between 1980 – 2000 and received 37 personal accounts of experiences of resistance. Resistance was found in two categories: skeptics towards a discovery that ultimately received a Nobel Prize and rejections of papers reporting a discovery or a contribution that was later awarded with a Nobel Prize. Campanario located passages from autobiographies and personal accounts to illustrate the nature of rejection or resistance. I added brief contextual information to the following examples for clarity.

The Nobel Prize in Physiology or Medicine 1958 was awarded jointly to George Wells Beadle and Edward Lawrie Tatum *"for their discovery that genes act by regulating definite chemical events"* and the other half to Joshua Lederberg *"for his discoveries concerning genetic recombination and the organization of the genetic material of bacteria"*. In his 1974 recollections, Beadle described that "... few people were ready to accept what seemed to us to be a compelling conclusion." (Beadle, 1974).

The Nobel Prize in Physics 1964 was awarded jointly to Charles Hard Townes, Nicolay Gennadiyevich Basov and Aleksandr Mikhailovich Prokhorov *"for fundamental work in the field of quantum electronics, which has led to the construction of oscillators and amplifiers based on the maser-laser principle"*. Townes recalled the pressure from peers who themselves were Nobel Laureates: *"One day...Raby and Kusch, the former and current chairmen of the department, both of them are Nobel laureates for their work with atomic and molecular beams and with a lot of weight behind their opinions, came into my office and sat down. They were worried. Their research depended on support from the same source as did mine. 'Look', they said, 'you should stop the work you are doing. You're wasting money. Just stop'."* (Lamb, Schleich, Scully, & Townes, 1999).

The second category of resistance identified by Campanario is rejections of Nobel class papers. He acknowledged that some of the rejections were indeed justified and in some cases the Nobel Prize winning versions differ from the initial versions.

Nobel laureate Allan Cormack's publications in the well-known Journal of Applied Physics in 1963 and 1964 were examples of so-called sleeping beauties, i.e. publications with delayed recognition. These articles introduced his theoretical underpinnings of computed tomography (CT) scanning. However, the two articles generated little interest — they attracted 7 citations alto-

[8] http://www2.uah.es/jmc/nobel/nobel.html

gether for the first 10 years — until Hounsfield used Cormack's theoretical calculations and built the first CT scanner in 1971. Cormack and Hounsfield shared the 1979 Nobel Prize in Physiology or Medicine.

Nobel laureate Stanley Prusiner wrote[9], *"while it is quite reasonable for scientists to be sceptical of new ideas that do not fit within the accepted realm of scientific knowledge, the best science often emerges from situations where results carefully obtained do not fit within the accepted paradigms."* Prusiner's comment echoes our earlier discussions on the potential biases of one's mindset. Both researchers and reviewers are subject to such biases. While it is reassuring to know that some early rejected works do become recognized later on, it is hard to find out how many, if any, potentially important discoveries discontinued because their values were not recognized soon enough.

3.3 Challenges to be Creative

The main thesis in the first half of the chapter is that human perception and cognition is biased. We are conservative and not particularly good at adopting a new perspective even at the presence of otherwise meaningful signs. The theme of the second half of the chapter is that our imagination is rather limited. Research has shown that the quality of hypotheses increases as more and more hypotheses are generated. However, analysts are more likely to select the first hypothesis that appears good enough than choose the best from all feasible options. We need extra assistances to stretch our imagination significantly.

3.3.1 Reasoning by Analogy

Aliens in Hollywood movies are often human-like creatures, computer transformed earth animals, or combinations of both. When college students were asked to create imaginary creatures they would expect to encounter on an alien planet, most of them re-combine and re-organize features found on earth animals. Our imagination is fundamentally biased by what we have seen and what we have experienced.

Reasoning by analogy is a frequently used strategy in science. The strategy duplicates the path of a previous success. One example is found in mass extinction research. An impact theory was originally proposed in 1980 to explain what caused the mass extinctions about 65 million years ago, known as the KT mass extinction. The extinction of dinosaurs was one of the consequences of the mass extinction. The impact theory suggested that an asteroid collided into the Earth and the dust of the impact covered the Earth's at-

[9] http://www.nobel.se

mosphere for a couple of years. In 1990, a big crater was discovered in the Mexico Bay and the crater was believed to be the most direct piece of evidence for the impact theory. In 2001, inspired by the success of the impact theory in explaining the KT extinction, researchers proposed a new line of research that would follow the same pattern of the impact theory's success. The difference was that it aimed to explain an even earlier mass extinction 250 million years ago. However, the validity of the analogy was questioned by more and more researchers. By 2010, it is the consensus that the analogy does not seem to hold given the available evidence. We were able to detect the analogical path from citation patterns in the relevant literature in our paper published in 2006 (Chen, 2006). The same conclusion was reached by domain experts in a 2010 review (French & Koeberl, 2010). We will revisit this example in later chapters of this book.

3.3.2 Competing Hypotheses

We are often torn by competing hypotheses. Each hypothesis on its own can be very convincing, while they apparently conflict with each other. One of the reasons we find hard to deal with such situations is because most of us cannot handle the cognitive load needed for actively processing multiple conflicting hypotheses simultaneously. We can focus on one hypothesis, one option, or one perspective at a time. In the literature, the commonly accepted magic number is 7, taken or given 2. In order words, if a problem involves about 5∼9 aspects, we can handle them fine. If we need to do more than that, we need to, so to speak, externalize the information, just as we need a calculator or a piece of paper to do multiplications beyond single digit numbers.

It is much easier to convince people using vivid, concrete, and personal information than using abstract, logical information. Even physicians, who are well qualified to understand the significance of statistical data, are convinced more easily by vivid personal experiences than by rigorous statistical data. Radiologists who examine lung x-rays everyday are found to have the lowest rate of smoking. Similarly, physicians who diagnosed and treated lung cancer patients are unlikely to smoke.

Analysis of Competing Hypotheses (ACH) is a procedure to assist the judgment on important issues in a complex situation. It is particularly designed to support decision making involving controversial issues by keeping track what issues analysts have considered and how they arrived at their judgment. In order words, ACH provides the provenance trail of a decision making process.

The ACH procedure has eight steps (Heuer, 1999):

1) Identify the possible hypotheses to be considered. Use a group of analysts with different perspectives to brainstorm the possibilities.
2) Make a list of significant evidence and arguments for and against each

hypothesis.

3) Prepare a matrix with hypotheses across the top and evidence on the side. Analyze the "diagnosticity" of the evidence and arguments — that is, identify which items are most helpful in judging the relative likelihood of the hypotheses.
4) Refine the matrix. Reconsider the hypotheses and delete evidence and arguments that have no diagnostic value.
5) Draw tentative conclusions about the relative likelihood of each hypothesis. Proceed by trying to disprove the hypotheses rather than prove them.
6) Analyze how sensitive your conclusion is to a few critical items of evidence. Consider the consequences for your analysis if that evidence were wrong, misleading, or subject to a different interpretation.
7) Report conclusions. Discuss the relative likelihood of all the hypotheses, not just the most likely one.
8) Identify milestones for future observation that may indicate events are taking a different course than expected.

The concept of diagnostic evidence is important. The presence of diagnostic evidence removes all the uncertainty in choosing one hypothesis over an alternative. Evidence that is consistent with all the hypotheses has no diagnostic value. Many illnesses may have the fever symptom, thus fever is not diagnostic evidence on its own. In the mass extinction example, the analogy of the KT mass extinction does not have sufficient diagnostic evidence to convince the scientific community. We use evidence in such situations to help us estimate the credibility of a hypothesis.

3.4 Boundary Objects

The concept of *boundary objects* is useful for understanding and externalizing communications involving different perspectives (Star, 1989; Bowker & Star, 2000; Wenger, 1998). Boundary objects are artifacts that are common between different groups of people and flexible enough that each group may find more room to develop. Boundary objects are externalized ideas. They form a shared context for communication. People can point to them in their communication as an explicit reference point. The real value of a boundary object is its potential to stimulate communications and exchanges of ideas that neither parties have thought of before hand.

We see what we expect to see. The same person may see different things in the same image at different times. Different people could see different things in the same image. Images taken by the Hubble Space Telescope mean different things to scientists and the public. It is a common practice that scientists make aesthetic enhancements or alterations to the original images before they are released to the public so that the pictures are easy to interpret. However, to astronomers, the "pretty pictures" that are intended for the public are qualitatively different from the purely "scientific images." To the public, these "pretty pictures" are taken as scientific rather than aesthetic. The *Pillars of Creation*[10] was a famous example of such public-friendly pictures. The public not only treated it as a scientific image, but also attached additional interpretations that were not found in the original scientific image (Greenberg, 2004).

The Eagle nebula is a huge cloud of interstellar gas in the south-east corner of the constellation Serpens. Jeff Hester and Paul Scowen at Arizona State University took images of the Eagle nebula using the Hubble Space Telescope. They were excited when they saw the image of three vertical columns of gas. Hester recalled, "we were just blown away when we saw them." Then their attention was directed to "a lot of really fascinating science in them" such as the "star being uncovered" and "material boiling" off of a cloud.

Greenberg (2004) described how the public reacted to the image. The public first saw the image on CNN evening news. CNN received calls from hundreds of viewers who claimed that they saw apparition of Jesus Christ in the Eagle nebula image. On CNN's live call-in program next day, according to viewers, they were able to see more: a cow, a cat, a dog, Juesus Christ, the Statue of Liberty, and Gene Shalit (a prominent film critic).

The reactions to the Eagle nebula image illustrate the concept of a boundary object, which is subject to reinterpretation by different communities. A boundary object is both vague enough to be adopted for local needs and yet robust enough to maintain a common identify across sites. Different communities, including astronomers, religious groups, and the public, put various meanings to the image. More interestingly non-scientific groups were able to make use of the scientific image and its unchallenged absolute authority for their own purposes so long as newly added meanings do not conflict with the original scientific meaning. Greenberg's analysis underlines that the more the scientific process is black-boxed, the easier it becomes to augment scientific knowledge with other extra-scientific meanings.

3.5 Early Warning Signs

It is usually easy to prove that something really exists, provided that we have the right equipments for detection and observation. It is almost impossible to prove that something does not exist. An example of the former is the discovery of an impact crater as the diagnostic evidence for the impact theory of mass extinctions 65 million years ago. An example of the latter is the failure of the intelligence to find evidence that Saddam did not have weapons of mass destruction.

The Black Swan is the bestseller of Nassim Nicholas Taleb. He was concerned with the influence of highly improbable and unpredictable events that

[10] http://apod.nasa.gov/apod/ap070218.html

have massive impact. He called rare, high-impact events Black Swans. Up to the 17th century, white swans were the only swans that had ever been seen in Europe. No one knew whether black swans actually existed until they were found in Australia. Taleb argued that while we focus too much on the odds that past events will repeat, which is the basis of many attempts to predict the future, the really important events are rare and unpredictable. The major myth of a black swan is that they are impossible to foresee. If all we have seen are white swans, can we conclude that all swans are white?

About 34 million years ago, after in a tropical state for many million years, the Earth changed suddenly from the tropical state to a colder state, a transition known as a greenhouse-icehouse transition. Epileptic seizures are a transient symptom often associated with a sudden contraction of a group of muscles. Sudden shifts from one state to another can be found in many complex dynamical systems. A key question is whether there are early warning signs before such sudden and radical changes occur.

Ecosystems, financial markets, and the climate are all complex dynamical systems. Asthma attacks and epileptic seizures are examples of spontaneous systemic failures. A sudden collapse of a civilization due to overpopulation is another type of example. This type of system-wide changes is often characterized as state transitions. It may be more likely for a system to change from one state to another. A global change may or may not be desirable. Some changes are not reversible. For instance, many are concerned whether the current climate change could lead to a possibly irreversible catastrophic change of the entire ecosystem. The concerns associated with the Gathering Storm boil down to the question whether economic and political changes may trigger crises that may bring down the entire system. Interestingly, the predictability of a financial market is something that the financial system would like to eliminate rather than enhance. On the other hand, one example of an early-warning signal is a measure of increased trade volatility. In contrast, scientific revolutions, as one would expect from high-risk and high-payoff transformative research, are intended and desirable.

Nature recently published a review article on early-warning signs for critical state transitions in complex dynamical systems (Scheffer, Bascompte, Brock, Brovkin, Carpenter, Dakos, Held, van Nes, Rietkerk, & Sugihara, 2009). The review article is written by 10 authors from 4 countries — Germany, The Netherlands, Spain, and U.S. — and multiple disciplines such as environmental sciences, economics, oceanography, and climate impact research.

Critical transitions can be explained in terms of bifurcations. A bifurcation means when a small smooth parameter change causes a sudden qualitative change in a system's behavior. Tipping points, butterfly effect, and phrase transition are all relevant concepts. Predicting such critical transitions in advance is extremely difficult because the system may show little change before the system suddenly shifts to a different state. Just like sighting white swans does not tell us anything about the arrival of a black swan.

A fundamental question concerning such critical transitions is whether the system gives any early signs as it approaches to such tipping points. The review in *Nature* found that although predicting such critical points in advance is extremely difficult, research in different scientific fields is now suggesting the existence of generic early warning signals. If, as these 10 authors concluded in their review, the dynamics of system near a critical point have generic properties regardless of differences in the details of each system, as the authors claimed in the paper, then it is indeed a profound finding.

The most important clues of whether a system is getting close to a critical threshold are related to a phenomenon called critical slowing down in dynamical system theory. To understand critical slowing down, we need to explain a few concepts such as fixed points, bifurcations, and fold bifurcations. A fixed point, also known as an invariant point of a function is a point that is mapped to itself by the function. As far as the fixed point is concerned, it remains unaffected by the function or mapping. Fixed points are used to describe the stability of a system. In a dynamical system, a bifurcation represents the sudden appearance of a qualitatively different solution for a nonlinear system as some parameters change. A bifurcation is a separation of a structure into two branches.

At fold bifurcation points (e.g., one stable and one unstable fixed points), the system becomes increasingly slow in recovering from perturbations. Research has shown that 1) such slowing down typically starts far from the bifurcation point, and 2) recovery rates decreases smoothly to zero as the critical point is approached. The change of the recovery rate provides important clues of how close a system is to a tipping point. In fact, the phenomenon of critical slowing down suggests three possible early-warning signals in the dynamics of a system approaching a radical change: slower recovery from perturbations, increased autocorrelation, and increased variance.

The authors of the review article stress that the work on early-warning signals in simple models is quite strong and it is expected that similar signals may arise in highly complex systems. They also note that more work is needed, especially in areas such as detecting patterns in real data and dealing with challenges associated with handling false positive and false negative signals. It is also possible that sudden shifts in a system may not necessarily follow a gradual approach to a threshold.

3.6 Summary

In this chapter, we have seen many common cognitive pitfalls that may undermine our creativity. In particular, these pitfalls may obscure our ability to detect early signs, to re-examine our existing mental model, to reduce possible biases in analytic reasoning, and to solve problems from diverse and possibly conflicting perspectives.

Mental models are valuable because they provide us a framework to develop an understanding of a situation or the state of a system. Mental models also make it possible to communicate complex phenomena. On the other hand, we are often reluctant to make fundamental changes to an established mental model. Not only can we resist evidence that challenges our mental model, but also turn such evidence around and strengthen our mental model. Mental models are like glasses: they help us to see the world in some ways, but also deprive us from see the world in alternative ways.

Many scientific breakthroughs and highly creative discoveries simply do not have any early signs that one can utilize or exploit in advance. Nevertheless, many predictive analytic systems are built on the assumption that what happened in the past will be repeated in the future. The questions concerning early warning signs and how to measure the transformative potential of research programs in their cradles are among the most challenging but crucial ones for science policy and research evaluation. Lessons learned from the 911 terrorist attacks and the puzzles of Iraqi WMDs have highlighted the nature of challenges when we face the unknown. The nature and the world around us have never stopped surprising us.

We may need more risk-taking reviewers to recognize the transformative potential of new research as well as to safeguard the integrity of science. More importantly, we need to realize that the diverse body of the literature seems to suggest that areas where the most creative work is likely to emerge or sparkle is where distinct and even conflicting views of the same phenomena run into one another. We need new ways of thinking and new tools that can augment our abilities to handle such situations more efficiently.

References

Beadle, G.W. (1974). Recollections. Annual Review of Genetics, 43, 1-13.

Ben-Zvi, A. (1976). Hindsight and foresight: A conceptual framework for the analysis of surprise attacks. World Politics, 28(3), 381-395.

Betts, R.K. (2007). Two faces of intelligence failure: September 11 and Iraq's missing WMD. Political Science Quarterly, 122(4), 585-606.

Bowker, G.C., & Star, S.L. (2000). Sorting things out — Classification and its consequences, Cambridge, MA.: MIT Press.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57(3), 359-377.

Drews, J. (2000). Drug discovery: A historical perspective. Science, 287(5460), 1960-1964.

French, B.M., & Koeberl, C. (2010). The convincing identification of terrestrial meteorite impact structures: What works, what doesn't, and why. Earth-Science Reviews, 98, 123-170.

Greenberg, J.M. (2004). Creating the "Pillars": Multiple meanings of a Hubble image. Public Understanding of Science, 13, 83-95.

Heuer, R.J. (1999). Psychology of intelligence analysis. Central Intelligence Agency.

Kneller, R. (2010). The importance of new companies for drug discovery: origins of a decade of new drugs. Nature Reviews Drug Discovery, 9, 867-882.

Lamb, W.E., Schleich, W.P., Scully, M.O., & Townes, C.H. (1999). Laser physics: Quantum controversy in action. Reviews of Modern Physics, 71, S263-S273.

Lipinski, C., & Hopkins, A. (2004). Navigating chemical space for biology and medicine. Nature, 432(7019), 855-861.

Perrow, C. (1984). Normal accidents: living with high-risk technologies, Princeton University Press.

Scheffer, M., Bascompte, J., Brock, W.A., Brovkin, V., Carpenter, S.R., Dakos, V., et al. (2009). Early-warning signals for critical transitions. Nature, 461(7260), 53-59.

Star, S.L. (1989). The structure of Ill-structured solutions: Boundary objects and heterogeneous distributed problem solving. In M. Huhs & L. Gasser (Eds.), Readings in Distributed Artificial Intelligence 3 (pp. 37-54). Menlo Park, CA: Kaufmann.

Sternitzke, C. (2010). Knowledge sources, patent protection, and commercialization of pharmaceutical innovations. Research Policy, 39(6), 810-821.

Ware, C. (2008). Visual thinking for design. Morgan Kaufmann.

Wenger, E. (1998). Communities of practice — learning, meaning, and identity. Cambridge, UK: Cambridge University Press.

Wohlstetter, R. (1962). Pearl Harbor: Warning and decisions:. Stanford University Press.

Chapter 4 Recognizing the Potential of Research

Basic research often does not have an earlier sign of whether and how it might be practically valuable. Research has found a recurring pattern that many scientific breakthroughs emerge as multiple lines of research converge. The question is: Is it possible to recognize a fruitful path ahead of time? In this chapter we discuss lessons learned from studies of both hindsight and foresight of identifying and recognizing the most important discoveries and innovations.

4.1 Hindsight

What can we learn from the past? How were scientific breakthroughs made?

4.1.1 Hibernating Bears

Black bears hibernate for 5∼7 months. When they wake up, they are as strong as ever. In contrast, if we are inactive for as short as several days, we may start to get weaker rather than stronger. We could start to lose our bone mass and strengths. People who are unable to maintain their usual levels of activity need to be very careful. For example, astronauts spend days in space need to have specially designed programs to keep themselves strong.

What makes the difference between human beings and black bears? This is the type of questions that everyone could see its value even before it gets answered. Seth Donahue and colleagues at Michigan Technological University started off with the good question. They were able to isolate a bone-building biomarker in the blood of black bears. The research has great commercial implications for osteoporosis treatment and prevention.

Their publication records show that their 2004 paper, entitled "Bending properties, porosity, and ash fraction of black bear (Ursus americanus) corti-

C. Chen, *Turning Points*

cal bone are not compromised with aging despite annual periods of disuse," was cited 13 times by 2010, whereas their 2006 paper, entitled "Hibernating bears as a model for preventing disuse osteoporosis," was cited 3 times. The practical value is more explicit in the 2006 paper. Donahue's technique was licensed to a company founded in 2007 called Aursos to make the therapeutic compounds for osteoporosis patients. Their story became one of the 100 successful stories in 2010 of how federal funding enables basic research and create jobs (The Science Coalition, 2010).

The connection between the basic research and its practical value is easy enough to spot in this case. The successful commercialization had made it easier for the funding agencies to justify their funding decisions when the research was in its cradle.

Scientists, social scientists and politicians frequently credit basic science with stimulating technological innovation, and with its economic growth. Despite a substantial body of research investigating this general relationship, relatively little empirical attention has been given to understanding the mechanisms that might generate this linkage. Researchers considered whether more rapid diffusion of knowledge, brought about by the norm of publication, might account for part of this effect (Sorenson & Fleming, 2004). They identify the importance of publication by comparing the patterns of citations from future patents to three groups of focal patents: (i) those that reference scientific (peer-reviewed) publications, (ii) those that reference commercial (non-scientific) publications; and (iii) those that reference neither. Their analyses strongly indicated publication as an important mechanism for accelerating the rate of technological innovation: Patents that reference published materials, whether peer-reviewed or not, receive more citations, primarily because their influence diffuses faster in time and space.

In parallel to the role of citation data in modeling and visualizing scientific revolutions, patent citation patterns play an important role in the construction of knowledge diffusion examples (Jaffe & Trajtenberg, 2002).

There are a number of extensively studied knowledge diffusion, or knowledge spillover, cases, namely liquid crystal display (LCD), nanotechnology (Braun, Schubert, & Zsindely, 1997; Meyer, 2000), and tissue engineering (Chen & Hicks, 2004).

In addition, knowledge diffusion between basic research and technological innovation (see Meyer, 2000; Narin & Olivastro, 1992), is also intrinsically related.

Empirical evidence shows a tendency of geographical localization in knowledge spillovers (Jaffe & Trajtenberg, 2002). Further studies have revealed profound implications of social dynamics. Agrawal, Cockburn and McHale (2003) show that social ties between collaborative inventors play a stronger part than geographic proximity in knowledge diffusion: inventors' patents are continuously cited by their colleagues in their former institutions.

Singh (2004) considered not just direct social ties but also indirect ones in social networks of inventors' teams based on data extracted from U.S.

Patent Office patents from 1975 – 1995. Two teams are connected in the social network if they have a common inventor. He used this network to analyze knowledge flows between teams based on patent citations among over half a million patents from 1986 – 1995. Social links between teams are associated with higher probability of knowledge flow. The probability decreases as social distance increases. An interesting finding in his study is that social links further explain why knowledge spillovers appear to be geographically localized. He also found a close social link to be a good predictor of knowledge flow regardless corresponding geographic proximity. In social network analysis, such networks of patents and inventors are known as an *affiliation network* (Wasserman & Faust, 1994). This affiliation network consists of two kinds of nodes: the inventors (the "actors"), and the patents (the "events").

LCD first appeared in 1968 and was subsequently improved several times between 1969 and 2003. Nanotechnology has various potential applications such as self-replicating nanobot and smart materials for artificial drugs and self-healing materials. Tissue engineering uses a combination of cells, engineering materials, and suitable biochemical factors to improve or replace biological functions in an effort to effect the advancement of medicine. Science and technology linkages are particularly valuable for funding agencies to evaluate funding efficiencies, for science policy researchers to study science and technology indicators, and even for investment fund managers to rank companies based on their innovation potentials.

4.1.2 Risks and Payoffs

A turning point in U.S. science policy was 1967 and 1968. Up until then, science policy had been dominated by the cold war. By 1963, the national investment in R&D was approaching 3% of GDP, 2.5 times the peak reached just before the end of World War II. More than 70% of this effort was supported by the federal government. 93% of it came from only three federal agencies: the Department of Defense (DOD), the Atomic Energy Commission (AEC), and the National Aeronautics and Space Administration (NASA). Their mission was to ensure the commercial spinoff of the knowledge and technologies they had developed. Much of the rest of federal R&D was in the biomedical field. Federal funding for basic research in universities reached a peak in 1967. It declined after that in real terms until about 1976.

The current funding environment is very competitive because of the tightened budget and increasing demands from researchers. In addition, the view that science needs to serve the society means that funding authorities as well as individual scientists need to justify how scientific inquiries meet the needs of society in terms of economic, welfare, and national security and competitiveness. There are two types of approaches to assess the quality and impact of scientific activities and identify priority areas for strategic planning. One is qualitative in nature, primarily based on opinions and judgments of experts, including experts in scientific fields, specialists in relevant areas of applications, and end users. The other is quantitative in nature, mainly including the development and use of quantitative metrics and indicators to provide evidence for making assessments and decisions. Metrics and indicators that can rank individual scientists, institutions, countries, as well as articles and journals have become increasingly popular. *Nature* recently published a group of featured articles and opinion papers on the issue of assessing assessment (Editorial, 2010).

Funding agencies have been criticized for their peer review systems for being too conservative and reluctant to support high-risk and unorthodox research. Chubin and Hackett (1990) found 60.8% of researchers supported this notion and 17.7% disagreed. Peer review has been an authoritative mechanism used to select and endorse research proposals and publications. Chubin and Hackett describe peer review as an intensively private process: it originates within a scientist's mind, continues on paper as a bureaucratic procedure, and ends behind the closed doors of a funding agency.

Laudel studied how researchers in Germany and Austrialia adapt their research in response to the lack of recurrent and external funding (Laudel, 2006). She also confirmed the perception that mainstream research is a key. One scientist said that he would not send a grant proposal unless he has at least 2 publications in the same area. She also noted the switch from basic research to applied research. Interdisciplinary research is also among the types of research that suffers from the tightened funding climate.

The profound problem in this context is closely related to the one that has been troubling the intelligence community. There is simply too much information and yet too few salient and meaningful signs. In addition, data cannot speak for itself. We need to have theories and mental models to interpret data and make sense of patterns emerging from data. There are many theories of how science evolves, especially from the philosophy of science, but these theories are so different that they explain the same historical event in science in totally different perspectives.

A considerable amount of inventions are built upon previously known technological features. Hsieh (2010) suggested an interesting perspective to see new inventions as a compromise between two contradictory factors: the usefulness of an invention and the relatedness of prior inventions that the new invention is to synthesize. He referred to such prior inventions as features. If these features are closely related, then the inventor will have to spend a lot of effort to distinguish them. On the other hand, if these features are barely related, the inventor will have to find out how they might be related. The optimal solution would be somewhere between the two extreme situations. The relations between features should neither be too strong nor too weak. The most cost-effective strategy for the inventor is to minimize the high costs of connecting unrelated features and, simultaneously minimize the costs of synthesizing ones that are already tightly connected. Hsieh tested his hypothesis with U.S. patents granted between 1975 and 1999. The usefulness of an invention was measured by future citation, while the relatedness of inventions was measured by the network of citations. He found a statistically significant inverse U-shaped relationship between an invention's usefulness and the relatedness among its component features. The usefulness of an invention was relatively low when the relatedness was either too strong or too weak. In contrast, the usefulness was the highest when the relatedness was in between the two extremes.

Transformative research is often characterized as being high risk and potentially high payoff. Revolutionary and groundbreaking discoveries are hard to come by. What are the implications of the trade-off strategy on funding transformative research with public funding? It is known that it takes long time before the values of scientific discoveries and technological innovations become clear. Nobel Prize winners are usually awarded for their work a few decades ago. We also know that Nobel class ideas do get rejected. The question is: to what extent will we be able to foresee scientific breakthroughs?

4.1.3 Project Hindsight

How long does it take for the society to fully recognize the value of scientific breakthroughs or technological innovations? *Project Hindsight* was commissioned by the U.S. Department of Defense (DoD) in order to search for lessons learned from the development of some of the most revolutionary weapon systems. One of the preliminary conclusions drawn from *Project Hindsight* was that basic research commonly found in universities didn't seem to matter very much in these highly creative developments. It appears, in contrast, that projects with specific objectives were much more fruitful.

In 1966, a preliminary report of *Project Hindsight* was published[1]. A team of scientists and engineers analyzed retrospectively how 20 important military weapons came along, including Polaris and Minuteman missiles, nuclear warheads, C-141 aircraft, and Mark 46 torpedo, and the M 102 Howitzer. Researchers identified 686 "research or exploratory development events" that were essential for the development of the weapons. Only 9% were regarded as "scientific research" and 0.3% was base research. 9% of research was conducted in universities.

Project Hindsight indicated that science and technology funds deliberately invested and managed for defense purposes have been about one order of magnitude more efficient in producing useful events than the same amount of funds invested without specific concerns for defense needs. *Project Hindsight* further concluded that:

1) The contributions of university research were minimal.

[1] Science, 1976, 192, pp. 105-111.

2) Scientists contributed most effectively when their effort was mission oriented.
3) The lag between initial discovery and final application was shortest when the scientist worked in areas targeted by his sponsor.

In terms of its implications on science policy, *Project Hindsight* emphasized mission-oriented research, contract research, and commission-initiated research. Although these conclusions were drawn from the study of military weapon development, some of these conclusions found their way to the evaluation of scientific fields such as biomedical research.

The extended use of preliminary findings had drawn considerable criticism. Comroe and Dripps (2002) criticized *Project Hindsight* as anecdotal and biased, especially because it was based on the judgments of a team of experts. In contrast to the panel-based approach taken by *Project Hindsight*, they started off with clinical advances since the early 1940's that have been directly responsible for diagnosing, preventing, or curing cardiovascular or pulmonary disease, stopping its progression, decreasing suffering, or prolonging useful life. They asked 40 physicians to list the advances they considered to be the most important for their patients. Physicians' responses were grouped into two lists in association with two diseases: a cardiovascular disease and a pulmonary disease. Then each of the lists was sent to 40∼50 specialists in the corresponding field. Specialists were asked to identify corresponding key articles, which need to meet the following criteria:

1) It had an important effect on the direction of subsequent research and development, which in turn proved to be important for clinical advance in one or more of the ten clinical advances they were studying.
2) It reported new data, new ways of looking at old data, new concept or hypothesis, a new method, new techniques that either was essential for full development of one or more of the clinical advances or greatly accelerated it.

A total of 529 key articles were identified in relation to 10 advances in biomedicine:

- Cardiac surgery
- Vascular surgery
- Hypertension
- Coronary insufficiency
- Cardiac resuscitation
- Oral diuretics
- Intensive care
- Antibiotics
- New diagnostic methods
- Poliomyelitis

It was found that 41% of these advances judged to be essential for later clinical advance were not clinically oriented at the time they were made. The scientists responsible for these key articles sought knowledge for the sake of knowledge. 61.7% of key articles described basic research, 21.2% reported other types of research, and 15.3% were concerned with development of new apparatus, techniques, operations, or procedures, and 1.8% were review articles or synthesis of the data of others.

Comroe and Dripps discussed research on research, similar to the notion of science of science. They pointed out that it requires long periods of time and long-term support to conduct and support retrospective and prospective research on the nature of scientific discovery and understand the courses of long and short lags between discovery and application. Their suggestion echoes the results of an earlier project commissioned by the NSF in response to *Project Hindsight*. NSF argued that the timeframe studied by the Hindsight project was too short to identify the basic research events that had contributed to technological advances. The NSF commissioned a project known as TRACES to find how long it would take for basic research to evolve to the point that potential applications become clear. However, Mowery and Rosenberg (1982) argued that the concept of research events is much too simplistic and the neither Hindsight nor TRACES used a methodology that is capable enough of showing what they purport to show.

4.1.4 TRACES

TRACES stands for *Technology in Retrospect and Critical Events in Science*. It was commissioned by the NSF. *Project Hindsight* looked back 20 years, but TRACES looked the history of five inventions and their origins dated back as early as 1850s. The five inventions are the contraception pill, matrix isolation, the video tape recorder, ferrites, and the electron microscope. TRACES identified 340 critical research events associated with these inventions and classified these events into three major categories: non-mission research, mission-oriented research, and development and application. 70% of the critical events were non-mission research, i.e. basic research. 20% were mission oriented, and 10% was development and application. Universities were responsible for 70% of non-mission and one third of mission oriented research. For most inventions, 75% of the critical events occurred before the conception of the ultimate inventions.

Hong Tseng, a program director at the NIH and an experienced user of CiteSpace, was the one who drew my attention to the TRACES. Funding agencies have many reasons to evaluate their research portfolios. They need to justify their funding decisions to the congress, the general public, and scientists. Strategically, there are two sets of requirements: the long term (10∼20 years), the short term (3∼5 years), and the near term (1∼2 years). The fast-increasing number of proposals received by many funding agencies on the one hand and the increasing recognition of the fundamental role of transformative and other high-risk, high pay-off research in sustaining the development of science and technology in a long run on the other hand pose a typical type of dilemma for decision making. In Chapter 5, we will discuss this issue from the perspective of optimal foraging in terms of maximizing the ratio of expected returns to risks.

The point of invention of video tape recorder was mid-1950s. It took almost 100 years to complete 75% of all relevant events, but the remaining 25% of the events converged rapidly within the final 10 years. In particular, the conception to innovation period took place in the final 5 years.

The invention of the video tape recorder involves 6 areas: control theory, magnetic and recording materials, magnetic theory, magnetic recording, electronics, and frequency modulation. The earliest non-mission research appeared in the area of magnetic theory. It was Weber's early ferromagnetic theory in 1852. The earliest mission-oriented research appeared in 1898 when Poulsew used steel wire for the first time for recording. According to TRACES, the technique was "readily available but had many basic limitations, including twisting and single track restrictions." Following Poulsew's work, Mix & Genest was able to develop steel tape with several tracks around 1900s, but limited by the lack of flexibility and increased weight. This line of invention continued as homogeneous plastic tape on the magnetophon tape recorder was first introduced in 1935 by AEG. A two layer tape was developed by 1940s. The development of reliable wideband tapes was intensive in early 1950s. The first commercial video tape record appeared in late 1950s.

The invention of electron microscope went through similar stages. The first 75% of research was reached before the point of invention and the translational period from conception to innovation.

The invention of electron microscope relied on five major areas, namely, cathode ray tube development, electron optics, electron sources, wave nature of electrons, and wave nature of light. Each area may trace several decades back to the initial non-mission discoveries. For instance, Maxwell's electromagnetic wave theory of light in 1864, Roentgen's discovery of emission of X-ray radiation in 1893, and Schrodinger's foundation of wave mechanics in 1926 all belong to non-mission research that ultimately led to the invention of electronic microscope. As a TRACES diagram shows, between 1860 and 1900 there was no connection across these areas of non-mission research. While the invention of electronic microscope was dominated by many earlier non-mission activities, the invention of video tape recorder revealed more diverse interactions among non-mission research, mission oriented research, and development activities.

Many insights revealed by TRACES have implications on today's discussions and policies concerning peer review and transformative research. Perhaps the most important lesson learned is the role of basic research, or non-mission research. As shown in the timeline diagrams of TRACES, an ultimate invention, at least in all the inventions studied by TRACES, emerged as multiple lines of research converged. Each line of research was often led by years and even decades of non-mission research, which was then in turn followed by mission-oriented research and development and application events.

In other words, it is evident that it is unlikely for non-mission research to foresee how their work will evolve and that it is even harder for non-mission research in one subfield to recognize potential connections with critical development in other subfields. Taken these factors together, we can start to appreciate the magnitude of the conceptual gulf that transformative research has to bridge.

4.2 Foresight

The term *foresight* refers to the systematic process of identifying strategic research areas and emerging generic technologies that are likely to yield the greatest economic and social benefits in the longer-term future of science, technology, and society (Anderson, 1997; Grupp & Linstone, 1999; Martin, 1995; Martin, 2010; Miles, 2010).

4.2.1 Looking Ahead

Since 1970, the Science and Technology Agency (STA) in Japan carried out a series of long-term forecasts, looking 30 years ahead into the future of science, technology, and innovation. These forecasts are one of the most systematic and wide-ranging forms of a foresight process. Some of the priority topics identified in forecasts made in early 1990s include the development of an HIV vaccine and effective methods for preventing Alzheimer's disease.

After 20 years passed since the first Delphi exercise, Japan's National Institute for Science and Technology Policy (NISTEP) reviewed the accuracy of its forecasts and found that 64% of topics were realized to some extent, but only 28% were fully realized. The accurate rate was overall regarded as encouraging given the experimental nature of the first Delphi exercise. In addition, the inaccurate results were more often due to political or social changes than technological development. Lessons learned from a separate analysis of the foresight indicated that expert panels used in such surveys should draw upon a wide range of expertise because experts tend to be over-optimistically biases about the development of their own fields. Interestingly, it was found that experts in neighboring fields were better able to foresee potential barriers in related topics. This finding underlines a central premise of this book: transformative discoveries are likely to emerge from the twilight zones where multiple fields meet.

Although it may be ironic that experts are more likely to be biased on topics that they have the most expertise, this is a valuable reminder of how vulnerable our cognitive abilities are. An interesting approach utilized by Japanese National Institute of Science and Technology Policy (NISTEP) is using science maps to depict hot research areas as mountains. Although the metaphor of landscape has been used in a variety of visualization for a long time, it is still rare to see such use in official reports of scientific priority forecasts. Hot research areas in NISTEP's science maps are defined as topic areas in which the total number of publications has exceeded a threshold.

NISTEP identified the promising role of a science map as a boundary object in facilitating communications between domain experts and facilitators: "During interviews, we were struck by the usefulness of the Science map as a basis for discussion . . . . With shared data such as the Science map, researchers from different fields can engage in more meaningful discussion of the development of scientific research. By sharing the same 'arena', researchers can mutually adjust their sense of distance, facilitating discussion among researchers or among researchers and policy makers. In the future, we would like to pursue this idea of the Science map as an arena for discussion." (Saka, Igami & Kuwahara., 2008).

It is certainly tempting for analysts to gather opinions from scientists about future development of scientific fields, but the question is how reliable the results are. The foresight approach in general is based on four principles (Martin, 1995):

1) The forecasts must incorporate economic and social needs;
2) It must cover all of science and technology;
3) It should evaluate the relative importance of different R&D tasks and determine priorities for policy purposes;
4) The forecast should be predictive (forecasting what is likely to happen) and normative (setting goals for what should happen).

Japan, the UK and Australia are widely known for their continued efforts in foresight processes. What has been done in the U.S. regarding the future of science and technology? During the 1960s, the Committee on Science and Public Policy (COSPUP) of the U.S. National Academy of Sciences conducted a series of *field surveys* in order to assess individual scientific disciplines and promising areas in these disciplines (Westheimer, 1965). The surveys were resumed in 1980 by the National Research Council (NRC). The Pimentel report *Opportunities in Chemistry* (Pimentel, 1985) was regarded as a successful field survey. A committee was set up in 1982 and chaired by Professor Pimentel with the goal to survey the research frontiers of chemistry. Several hundred chemists were asked to identify topics for further reviews.

Funding agencies and the U.S. congress criticized field surveys for several reasons. In particular, almost all the field surveys in the 1980s made demands unrealistically that funding for the field in question needs to be doubled over the next 5 years. Each field study on average cost \$0.5 million $\sim$ \$1.0 million and takes 3 years to complete. The final reports were often too long and inaccessible to outsiders. No attempt was made to identify any overall priorities. Field studies relied on informed but subjective judgments of experts. More importantly, field studies did not identify priority areas needed for science policy decision making in response to the stretched public funding. In part, the reluctance of identifying areas of declining importance was from the scientific community and the National Academies that serve the interest of the scientific community. Subsequent foresight activities learned from these lessons and placed more emphases on balancing between interested parties and independent third parties. Ben Martin's reviews (Martin, 1995; Martin, 2010) provide informative accounts of the history of foresight, including Australia, Germany, New Zealand, the Netherlands, and the UK.

4.2.2 Identifying Priorities

Goodwin and Wright (2010) reviewed forecasting methods that target for rare and high-impact events. They identified the following six types of problems that may undermine the performance of forecasting methods:

- Sparsity of reference class
- Reference class that is outdated or does not contain extreme events
- Use of inappropriate statistical models
- The danger of misplaced causality
- Cognitive biases
- Frame blindness

Goodwin and Wright identified three heuristics that can lead to systematically biased judgments: (1) availability, (2) representativeness, and (3) anchoring and insufficient adjustment. The availability heuristic bias means that human beings find easier to recall some events than others, but it usually does not mean that easy-to-recall events have a higher probability of occurring than hard-to-recall events. The representativeness heuristic is a tendency to ignore base-rate frequencies. The anchoring and insufficient adjustment means that forecasters make insufficient adjustment for the future conditions because they anchor on the current value. As we can see, cognitive biases illustrate how vulnerable human cognition is in terms of estimating probabilities intuitively. Expert judgment is likely to be influenced by these cognitive biases. Researchers have argued that in many real world tasks, apparent expertise may have little to do with any real judgment skills at the task in question.

The increasing emphasis on accountability for science and science policy is influenced by many factors, but two of them are particularly influential and persistent despite the fact that they started to emerge more than a decade ago. The two factors are 1) limited public funding, and 2) the growing view that publicly funded research should contribute to the needs of society (MacLean, Anderson, & Martin, 1998). The notion of a value-added chain is useful for explaining the implications. The earlier value-added chain, especially between 1940s and 1960s, was simple. Researchers and end-users were loosely coupled in such value-added chains. The primary role of researchers was seen as producing knowledge and the primary role of end-users was to make use of produced knowledge whenever applicable, passively. Decisions on how to allocate research funds were largely made based on the outcome of peer reviews. The peer here was the peer of scientists, but the peer of end-users was not part of the game. The focus was clearly and often exclusively on science.

The view that science should serve the needs of society implies that the linkage between science and end-users becomes an integral part of science policy and strategic, long-term planning, including identifying funding priorities and assessing the impact of research. As a result, the new value-added chain includes intermediate users as well as scientists and end-users. For example, following (MacLean et al., 1998), a simple value-added chain may include researchers in atmospheric chemistry, intermediate users from meteorological office and consultancy firms, and a supermarket chain as an end-user. Several ways are suggested to understand users, including their long-term or short-term needs, generic versus specific needs, proactive compared to reactive users, and end-users versus intermediate users. The role of intermediate users is to transform the scientific and technological knowledge and add value to such knowledge for the benefit of the following user in the chain. In one example of a complex value-added chain given in (MacLean et al., 1998), the value-added chain consists of three categories of stakeholders: researchers, intermediate users, and end-users. Aquatic pollution researchers may communicate with intermediate users such as sensor development companies, informatics companies, and pollution regulatory authorities. Intermediate users may have their own communication channels among themselves. Intermediate users in turn connect to water companies and polluting industry.

In soliciting users' opinions, especially from scientists, users tend to concentrate on shorter-term issues and more immediate problems than a 10-20 year strategic timeframe. One way to encourage users to articulate their longer-term research needs is to ask users a set of open-ended questions. Here are some examples of open-ended questions on long-term environmental research priority issues (MacLean et al., 1998):

- If someone can tell you precisely how environmental issues would affect your business in 10-20 years, what questions would you most wish to ask?
- What understanding about the environment do you not have at present, but would need in the next 5-20 years to enhance your organization's business prospects?
- If all the constraints, financial or otherwise, can be removed, what would you suggest that funding agencies could do in relation to environmental research in addition to what you have already mentioned?

In assessing science and technology foresights, one way to identify priorities is to solicit assessments from experts and users and organize their assessments along two dimensions: feasibility and attractiveness. The dialog between researchers and users is increasingly regarded as a necessary and effective approach to identify science and technology priorities. Scientists and researchers are likely to provide a sound judgment on what is feasible, whereas users often make a valuable input on what is attractive. This type of method was adopted by (MacLean et al., 1998) to identify the nature of links in a value-added chain and map science outputs on to user needs. Specially, a two-round Delphi survey was conducted. Responses from over 100 individuals were obtained in each round. The differences between responses in the two rounds were plotted in a two-dimensional feasibility-by-attractiveness space.

Fig. 4.1 shows a schematic illustration of the movements of assessments in this two-dimensional space, which is a representation of the linkage between scientists and users in the value-added chain model. For example, the topic *remote data acquisition* in the high feasibility and high attractiveness quadrant moved to a position with an even higher feasibility and a higher attractiveness after the second round of the Delphi survey. In contrast, the topic *sustainable use of marine resources* was reduced in terms of both feasibility and attractiveness. It is possible that one group of stakeholders changed their assessments, but the other group's assessments remained unchanged. For example, while users did not alter their attractiveness assessments of topics such as management of freshwater resources and prediction of extreme atmospheric events, scientists updated the corresponding feasibility assessments: one went up and the other went down.

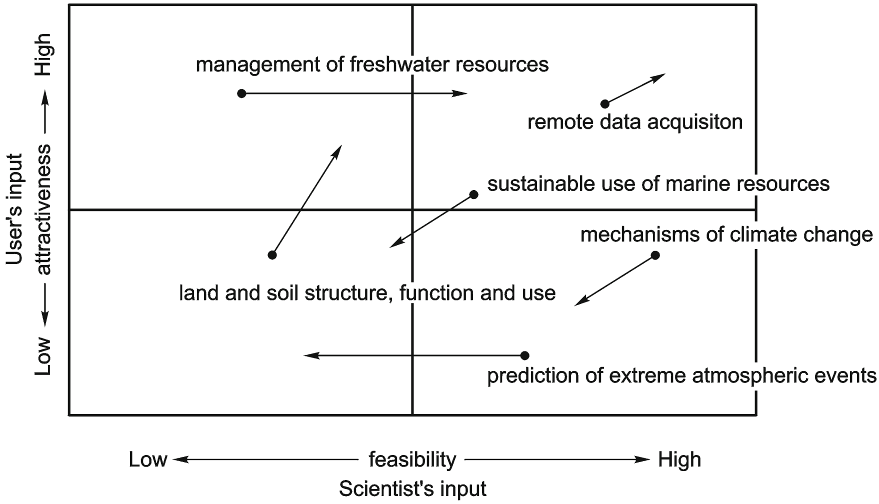

management of freshwater resources
remote data acquisiton
sustainable use of marine resources
mechanisms of climate change
land and soil structure, function and use
prediction of extreme atmospheric events
User's input
Low ← attractiveness → High
Low ← feasibility → High
Scientist's input

Fig. 4.1 Feasibility and attractiveness of research topics. Source: The diagram is drawn based on Figure 3 of (MacLean et al., 1998).

In contrast to the broadened social contract view of science in today's society, it is worth noting that there is a profound belief in the value of intellectual freedom and the serendipitous nature of science. Both sides have many tough questions to answer. Is there sufficient evidence based on longitudinal, retrospective, and comparative assessments of both priority areas chosen by science and technology foresight approaches and scientific and technological breakthroughs emerged and materialized regardless? Given the evidently changing consensus between different voting rounds of Delphi surveys, what are the factors that trigger the shift?

4.2.3 The Delphi Method

The Delphi method is the most frequently used method in foresight activities.

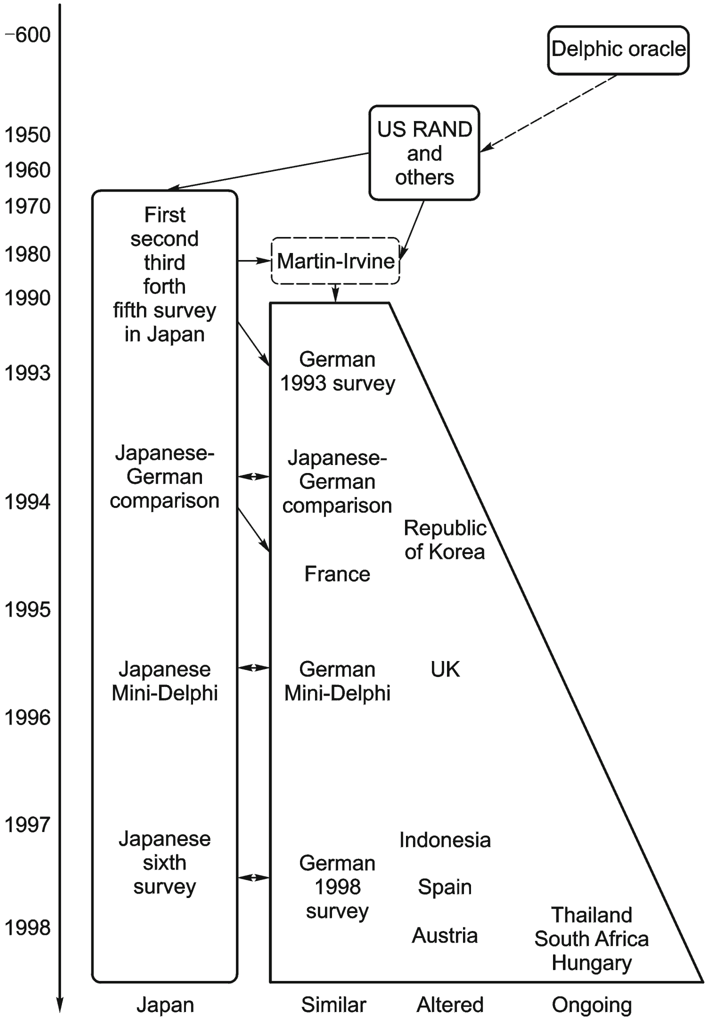


Fig. 4.2 A genealogical tree of national applications of the Delphi method. Source: Figure 1 in (Grupp & Linstone, 1999).

Some of the earliest studies using Delphi were performed at the RAND Corporation (Dalkey, 1969; Kaplan, Skogstad, & Girshick, 1950). In 1972, the Science and Technology Agency in Japan selected the Delphi method for foresight activities. It gathers experts' judgments using successive iterations of a survey questionnaire. Each iteration is also known as a *round*. The results of earlier rounds are shared among experts in a new round so that facilitators of the survey can identify the convergence of opinions or the persistence of different opinions (Grupp & Linstone, 1999). The Delphi method is considered particularly useful for making long-range forecasts over 20∼30 years because in such situations expert opinions are the only source of information available. Unlike a committee, which usually seeks consensus, Delphi does not force consensus. The Delphi method allows experts to shift their opinions from one round to another based on new information that becomes available to them. Fig. 4.2 shows a genealogical tree of national applications of the Delphi method. The diagram is adopted from (Grupp & Linstone, 1999).

A widely cited retrospective review of Japan's Delphi experiences was done by Cuhls (1998) in her doctoral dissertation. She found that the Japanese Delphi studies were able to keep the unresolved issues such as early earthquake detection on the national science and technology agenda even in years of no earthquakes when the public and policy makers paid little attention to these issues.

4.2.4 Hindsight on Foresight

How realistic and reliable are expert opinions obtained from foresight activities? There is a rich body of literature on Delphi and related issues such as individual opinion change and judgmental accuracy in Delphi-like groups (Rowe, Wright, & McColl, 2005), pitfalls and neglected aspects (Geels & Smit, 2000).

In a recent study, Felix Brandes (2009) addressed this issue and assessed expert anticipations in the UK technology foresight program. The UK's technology foresight program was recommended in the famous 1993 White Paper 'Realizing our Potential.' 15 expert panels were formed along with a large scale national Delphi survey. The survey was sent to 8,384 experts in 1994 to generate forecasts on 2015 and beyond. 2,585 responded to the survey. About 2/3 of statements were predicted to be realized between 1995 and 2004. Brandes' study was therefore to assess how realistic the 1994 expert estimates by 2004, i.e. 10 years later.

Out of the original 15 panels of the 1994 UK foresight program, Brandes selected three panels, Chemicals, Energy, and Retial & Distribution, to follow up the status of their forecasts in terms of *Realized*, *Partially Realized*, *Not Realized*, and *Don't Know*. An online survey was used and the overall response rate was 38%.

Brandes' *Hindsight on Foresight*[2] survey found only 5% of the Chemicals statements and 6% of the Retail & Distribution statements were regarded as realized by the experts surveyed, whereas 15% of Energy topics were realized. If the assessment criteria were relaxed to lump together fully and partially realized topics, known as the joint realization rate, Chemicals scored 28%, Energy 34%, and Retail & Distribution 43%.

In summary, the 1994 expert estimates were overly optimistic, which is a well-documented issue in the literature. Researchers have found that top experts tend to be even more optimistic than the overall response group (Tichy, 2004). According to (Tichy, 2004), the assessments of top experts tend to suffer from an optimism bias due to the experts' involvement and their underestimation of realization and diffusion problems. Experts working in business have a stronger optimism bias than those working in the academia or in the administration. Reasons that cause such biases are still not completely clear. More importantly, retrospective assessments of the status of priority topics identified by foresight activities are limited in that they do not provide overall assessments of topics and breakthroughs that were totally missed and unanticipated by foresight activities.

4.3 Summary

Many scientific breakthroughs and highly creative discoveries simply do not have any early signs that one can utilize or exploit in advance. On the other hand, many predictive analytic systems are built on the assumption that what happened in the past will be repeated in the future. The questions concerning early warning signs and how to measure the transformative potential of research programs in their cradles are among the most challenging but crucial issues for science policy and research evaluation. Lessons learned from TRACES highlight the role of non-mission research in inventions and the lack of early signs for practical potentials later on. Assessments of the priority areas identified by foresight activities pointed out the difficulty of experts in realistically judging feasibility of technical development but emphasized the useful interaction along a value-added chain of stakeholders.

We may need more risk-taking reviewers to recognize the transformative potential of new research as well as to safeguard the integrity of science. More importantly, we need to realize that the diverse body of the literature seems to suggest that areas where the most creative work is likely to emerge or sparkle is where distinct and even conflicting views of the same phenomena run into one another. We need new ways of thinking and new tools that can augment our abilities to handle such situations more efficiently!

[2] The Office of Science and Technology (UK) sent out a 'Hindsight on Foresight' survey in 1995.

References

Agrawal, A., Cockburn, I., & McHale, J. (2003). Gone but not forgotten: Labor flows, knowledge spillovers, and enduring social capital. NBER Working Paper No. 9950.

Anderson, J. (1997). Technology foresight for competitive advantage. Long Range Planning, 30(5), 665-677.

Brandes, F. (2009). The UK technology foresight programme: An assessment of expert estimates. Technological Forecasting and Social Change, 76(7), 869-879.

Braun, T., Schubert, A., & Zsindely, S. (1997). Nanoscience and nanotechnology on the balance. Scientometrics, 38, 321-325.

Chen, C., & Hicks, D. (2004). Tracking knowledge diffusion. Scientometrics, 59(2), 199-211.

Chubin, D.E., & Hackett, E.J. (1990). Paperless science: Peer review and U.S. science policy.

Comroe, J.H., & Dripps, R.D. (2002). Scientific basis for the support of biomedical science. In R.E. Bulger, E. Heitman & S.J. Reiser (Eds.), The ethical dimensions of the biological and health sciences (2nd ed., pp. 327-340). Cambridge, UK: Cambridge University Press.

Cuhls, K. (1998). Technikvorausschau in Japan. Heidelberg: Physica-Springer.

Dalkey, N.C. (1969). The Delphi method: An experimental study of group opinion. Santa Monica, CA: The Rand Corporation.

Editorial. (2010). Assessing assessment. Nature, 465, 845.

Geels, F.W., & Smit, W.A. (2000). Failed technology futures: Pitfalls and lessons from a historical survey. Futures, 32(9-10), 867-885.

Goodwin, P., & Wright, G. (2010). The limits of forecasting methods in anticipating rare events. Technological Forecasting and Social Change, 77(3), 355-368.

Grupp, H., & Linstone, H.A. (1999). National technology foresight activities around the globe — Resurrection and new paradigms. Technological Forecasting and Social Change, 60(1), 85-94.

Hsieh, C. (2010). Explicitly searching for useful inventions: Dynamic relaatedness and the costs of connecting versus synthesizning. Scientometrics.

Illinois Institute of Technology. (1969). Technology in retrospect and critical events in science. Chicago: The Illinois Institute of Technology Research Institute.

Jaffe, A., & Trajtenberg, M. (2002). Patents, citations & innovations. The MIT Press.

Kaplan, A., Skogstad, A.L., & Girshick, M.A. (1950). The prediction of social and technological events. Public Opinion Quarterly XIV, 93-110.

Laudel, G. (2006). The art of getting funded: How scientists adapt to their funding conditions. Science and Public Policy, 33(7), 489-504.

MacLean, M., Anderson, J., & Martin, B.R. (1998). Identifying research priorities in public sector funding agencies: Mapping science outputs on to user needs. Technology Analysis & Strategic Management, 10(2), 139-155.

Martin, B.R. (1995). Foresight in science and technology. Technology Analysis & Strategic Management, 7(2), 139-168.

Martin, B.R. (2010). The origins of the concept of 'foresight' in science and technology: An insider's perspective. Technological Forecasting and Social Change, 77(9), 1438-1447.

Meyer, M. (2000). What is special about patent citations? Differences between scientific and patent citations. Scientometrics, 49(1), 93-123.

Miles, I. (2010). The development of technology foresight: A review. Technological Forecasting and Social Change, 77(9), 1448-1456.

Narin, F., & Olivastro, D. (1992). Linkage between technology and science. Research

Policy, 21, 237-249.

Pimentel, G.C. (1985). Opportunities in Chemistry. Washington, DC: National Academy Press.

Rowe, G., Wright, G., & McColl, A. (2005). Judgment change during Delphi-like procedures: The role of majority influence, expertise, and confidence. Technological Forecasting and Social Change, 72(4), 377-399.

Saka, A., Igami, M., & Kuwahara, T. (2008). Science map 2006: study on hot research areas (2001 – 2006 by Bibliometric method). National Institute of Science and Technology Policy (NISTEP).

Singh, J. (2004, January 9). Social networks as determinants of knowledge diffusion patterns. Retrieved March 24, 2004, from http://www.people.hbs.edu/jsingh/academic/jasjit_singh_networks.pdf

Sorenson, O., & Fleming, L. (2004). Science and the diffusion of knowledge. Research Policy, 33(10), 1615-1634.

The Science Coalition. (2010). Sparking economic growth: How federally funded university research creates innovation, new companies, and jobs. The Science Coalition.

Tichy, G. (2004). The over-optimism among experts in assessment and foresight. Technological Forecasting and Social Change, 71(4), 341-363.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications. Cambridge University Press.

Westheimer, F.H. (1965). Chemistry: Opportunities and needs. Washington, DC: National Academy of Sciences.

Chapter 5 Foraging

Creative thinking in a broad range of scientific discoveries is similar to food foraging. A food forager uses creative processes when finding the next patch of food. Decisions made for optimal foraging need to take into account the uncertainties and risks of the investment of time, energy, and other resources and the expected gains. If foragers have a vast number of alternatives to consider but only a tiny chance of finding anything useful, then the foragers are alert to scents, signs, and other types of cues to avoid an unproductive search. A scientist, as a forager and creator of new knowledge, faces similar challenges of finding patches of ideas, theories, and evidence in scientific inquires. Since scientific breakthroughs, or transformative discoveries, are truly novel in creating a new way of thinking, these involve the identification of patches of knowledge that are often either remote from the state of the art or non-existent. Notable examples of this include searching for earth-like planets in the Universe, searching for satisfactory compounds in chemical space for drug discovery, or searching for new ideas that may revolutionize a field or lead to the birth of a new field.

Optimal foraging theory provides a surprisingly profound foundation for the study of searching for food, information, and ideas. It formulates the nature of such activities as a series of decisions to be made with the intent to maximize the ratio of reward to cost or overhead. The type of rewards could be food for food foragers, information for information searchers, evidence for intelligence analysts, or inspirational ideas for scientists. The type of cost and overhead includes the energy consumed for chasing a pray, time spent on search, and suspicious patterns. The optimal foraging perspective leads to a theory of discovery that consists of generic mechanisms of how creative discoveries are made. Our central hypothesis is that the linking of previously unconnected or loosely connected bodies of knowledge is a profound mechanism of scientific creativity. In addition, we expect that transformative advances are recognizable and detectable in terms of how they alter the existing structure of knowledge.

C. Chen, *Turning Points*

5.1 An Information-Theoretic View of Visual Analytics

The investigation of 911 terrorist attacks has raised questions on whether the intelligence agencies could have connected the dots and prevented the terrorist attacks (Anderson, Schum, & Twining, 2005). Prior to the September-11 terrorist attacks, several foreign nationals enrolled in different civilian flying schools to learn how to fly large commercial aircraft. They were interested in learning how to navigate civilian airlines, but not in landings or takeoffs. And they all paid cash for their lessons. What is needed for someone to connect these seemingly isolated dots and reveal the hidden story?

In an intriguing *The New Yorker* article, Gladwell differentiated puzzles and mysteries with the stories of the collapse of Enron (Gladwell, 2007). To solve the puzzle, more specific information is needed. To solve a mystery, one needs to ask the right question. Connecting the dots is more of a mystery than a puzzle. Solving mysteries is one of the many challenges for visual analytic reasoning. We may have all the necessary information in front of us and yet fail to see the connection or recognize an emergent pattern. Asking the right question is critical to stay on track.

In many types of investigations, seeking answers is only part of the game. It is essential to augment the ability of analysts and decision makers to analyze and assimilate complex situations and reach informed decisions. We consider a generic framework for visual analytics based on information theory and related analytic strategies and techniques. The potential of this framework to facilitate analytical reasoning is illustrated through several examples from this consistent perspective.

In information theory, the value of information carried by a message is the difference of information entropy *before* and *after* the receipt of the message. Information entropy is a macroscopic measure of uncertainty defined on a frequency or probability distribution. A key function of an information-theoretical approach is to quantify discrepancies of the information content of distributions. Information indices, such as the widely known Kullback-Leibler divergence (Kullback & Leibler, 1951), are entropy-based measures of discrepancies between distributions (Soofi & Retzer, 2002).

The Kullback-Leibler (K-L) divergence of probability distribution Q from probability distribution P is defined as follows:

$$\text{Divergence}_{K-L}(P : Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

The divergence measures the loss of information if Q is used instead of P, assuming P is the true distribution. Information entropy can be seen as the divergence from the uniform distribution. This is consistent with the common interpretation of information entropy as a measure of uncertainty, or the lack of uniformity.

A useful alternative interpretation of the K-L divergence is the expected extra message length to be entailed in communication if the message is transmitted without using a coding scheme based on the true distribution. In computer science, the Kolmogorov complexity of an object, also known as algorithmic entropy or program-size complexity, measures the amount of computational resources required to specify the object.

Integrating and interpreting findings at both microscopic and macroscopic levels are essential for visual analytics. Human analysts are good at dealing with macroscopic patterns and mysteries, but they would be overwhelmed if they were to deal with microscopic attributes and relations. Information-theoretic strategies and techniques are particularly valuable to visual analytics because they lend themselves to a wide variety of ways to decompose and aggregate information across microscopic and macroscopic levels of abstraction, for example, through statistical decomposition analysis. Information-theoretic analysis provides a unifying foundation of an integrative study of uncertainty, belief, and evidence of complex situations.

The following set of examples is discussed from the generic information-theoretic perspective. Although they are not necessarily visual analytic problems per se, they represent important challenges to analytical reasoning. Our intention is to stimulate further investigations of how information-theoretic strategies and techniques may draw our attention to informational patterns that may lead to potentially significant questions.

5.1.1 Information Foraging and Sensemaking

Sensemaking is a critical part of analytical reasoning. Sensemaking itself typically involves two integral and iterative sub-processes: information foraging and sensemaking. The following example illustrates how we can characterize information foraging and sensemaking from an information theoretical point of view.

Information foraging theory is a predictive model of information foragers' search behavior (Pirolli, 2007). According to the theory, an information environment is made of patches of information, and an information forager moves from one patch to another as they search for information, just as a predator looking for its prey in the animal world. The theory is developed to answer questions about how information foragers would choose patches to work with and what influence their decisions on how long they should spend their time with patches.

Information foraging theory is built on the assumption that people adapt their search strategies to maximize their profitability, or the profit-investment ratio. People may adapt their search by reconfiguring the information environment. The investment typically includes the time spent in searching and assimilating information in patches of information. The profit includes the gain by finding relevant information. Users, or information foragers, tend to follow a path that can maximize the overall profitability. *Information scent* is the perception of the value, cost, or accessible path of information sources. When possible, one relies on information scent to estimate the potential profitability of a patch.

The power of information foraging theory is its own adaptability and extensibility. It provides a quantitative framework for interpreting behavioral patterns at both microscopic and macroscopic levels. For instance, connecting the dots of mysterious behaviors of 911 hijackers at flying schools would depend on the prevalence and strength of the relevant information scent (Anderson et al., 2005). The question is where an analyst could draw the right information scent in the first place. Fig. 5.1 is not designed with information foraging theory in mind, but the visualization intuitively illustrates the profit maximization principle behind the theory. The connective density reinforces the boundaries of patches. Colors and shapes give various information scents about each patch in terms of its average age and the popularity of citations. These scents will help users to choose which patch they want to explore in more detail.

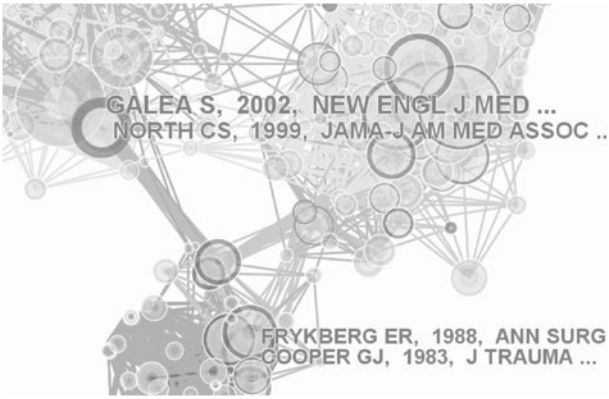

GALEA S, 2002, NEW ENGL J MED ...
NORTH CS, 1999, JAMA-J AM MED ASSOC ..
FRYKBERG ER, 1988, ANN SURG
COOPER GJ, 1983, J TRAUMA ...

Fig. 5.1 The three clusters of co-cited papers can be seen as three patches of information. All three patches are about terrorism research. Prominently labeled papers in each patch offer information scent of the patch. Colors of patches, indicating the time of a connection, provide a scent of freshness. The sizes of citation rings provide a scent of citation popularity. Source: (Chen, 2008). (see color figure at the end of this book)

From the information-theoretic view, information scent only makes sense if it is connected to the broader context of information foraging, including the goal of search, the prior knowledge of the analyst or the information forager, and the contextual situation. This implies a deeper connection between the information-theoretic view and various analytic tasks in sensemaking. The following is a sensemaking example in which an information-theoretic approach is applied to the study of uncertainty and influential factors involved in political elections.

Voting in political elections involves a complex sensemaking and reasoning process. Voters need to make sense overwhelmingly diverse information, differentiate political positions, accommodate conflicting views, adapt beliefs in light of new evidence, and make macroscopic decisions. Information-theoretic approaches provide a valuable and generic strategy for addressing these issues. Voters in political elections are influenced by candidates' positions regarding a spectrum of political issues and their own interpretations of candidates' positions (Gill, 2005).

Researchers are particularly interested in the impact of uncertainty concerning political positions of candidates from a voter's point of view. From an information-theoretic perspective, candidate positions on a variety of controversial issues can be represented as a probability distribution. The underlying true distribution is unknown. The voters' uncertainty can be measured by the divergence of a sample from the true distribution. In a study of the 1980 presidential election, Bartels finds that voters in general dislike uncertainty (Bartels, 1988).

In a study of a 1994 congressional election, Gill (2005) constructs an aggregate measure of uncertainty of candidates as well as political issues based on voters' self-reported 3-level certainty information regarding to each political issue question. The study analyzed answers from 1,795 respondents on several currently salient issues, such as crime, government spending, and healthcare. The results suggest that politicians would be better off with unambiguous positions, provided that those positions do not drastically differ from those held by widely supported candidates.

The effect of uncertainty seems to act at aggregated levels as well as individual levels. Crime, for example, has been a Republican campaign issue for decades. The study found that Republican candidates who were vague on this issue were almost certainly punished (Gill, 2005). This example shows that an information-theoretical approach provides a flexible tool for studying information uncertainty involved in complex reasoning and decision making processes.

5.1.2 Evidence and Beliefs

We review our beliefs when new information becomes available. For example, physicians run various tests with their patents. Physicians make sense of test results and decide whether more tests are needed. In general elections, voters ask questions about candidates' political positions in order to reduce or eliminate uncertainties about choosing candidates. Bayesian reasoning is a widely used method to analyze evidence and synthesize our beliefs. It has been used in a wide variety of application domains, from interpreting women's mammography for breast cancer risks to differentiating spam from genuine emails.

The search for the USS Scorpion nuclear submarine is a frequently told story of a successful application of Bayesian reasoning. The USS Scorpion was lost from the sea in May 1968. An extensive search failed to locate the vessel. The search was particularly challenging because of the lack of knowledge of its location prior to its disappearance. The subsequent search was guided by Bayesian search theory, which takes the following steps:

1) Formulate hypotheses of whereabouts of a lost object.
2) Construct a probability distribution over a grid of areas based on the hypotheses.
3) Construct a probability distribution of finding the lost object at a location if it is indeed there.
4) Combine the two distributions and form a probability distribution and use the new distribution to guide the search.
5) Start the search from the area with the highest probability and move to areas with the next highest probabilities.
6) Revise the probability distribution using the Bayesian theorem as the search goes on.

In the Scorpion search, experienced submarine commanders were called in to come up with hypotheses independently of whereabouts of the Scorpion. The search started from the grid square of the sea with the highest probability and moved on to squares with the next highest probabilities. The probability distribution over the grid was updated as the search moved along using Bayesian theorem. The Scorpion was found in October more than 10,000 feet under water within 200 feet of the location suggested by the Bayesian search.

The Bayesian method enables searchers to estimate the cost of a search at local levels and allows the searchers adapt their search path according to the revised beliefs as the process progresses. This adaptive strategy is strikingly similar to the profit maximization assumption of information foraging theory. The revision of our beliefs turns probabilistic distributions to information scents. The Bayesian search method is a tool that may help analysts in solving mysteries.

If solving mysteries in visual analytics is akin to finding needles in numerous haystacks, the needle of interest in visual analytics often has a low key or low profile. They tend to blend in well with others. Furthermore, human analysts are superior when it comes to identify and differentiate information that only has subtle differences from others. In order to find connections between a few low-profile needles, analysts need tools that can reliably single out subtle outliers or surprises from an overwhelmingly vast and diverse population. Information indices are designed to capture discrepancies of different distributions. The following example shows how such information indices are used to detect surprising spots in video frames.

5.1.3 Salience and Novelty

Salience and novelty of information are essential properties to visual analytics. A salient feature or pattern is prominent in the sense that it stands out perceptually, conceptually, and/or semantically. In contrast, novelty characterizes the uniqueness of information. A landmark has a high saliency in its skyline, whereas the novelty of a design is how unique it is in comparison to others. It is often effortless for humans to spot salient or novel features visually. However, it is a real challenge to identify these features computationally because these are emergent macroscopic features in nature as opposed to specific microscopic ones. To visual analytics, a fundamental challenge is to capture such emergent features and match semantic features with visual saliency and novelty to facilitate analytical reasoning.

From an information-theoretic point of view, saliency can be defined as statistical outliers in a semantic and/or visual feature space. Novelty, on the other hand, can be defined as statistical outliers along a specific dimension of the space, such as the temporal dimension.

A computational model developed by Itti and Paldi can detect surprising events in video based on the concepts of saliency and novelty (Itti & Baldi, 2005). The question of finding surprising scenes is formulated in terms of the piece of data that is responsible for how one's belief changes between two distinct frames. One's belief is transformed from a prior distribution $\boldsymbol{P}$(Model) to a posterior distribution $\boldsymbol{P}$(Model | Data). The difference between prior and posterior distributions over all models is measured by relative entropy, i.e. the K-L divergence. Surprise is defined as the average of the log-odd ratio with respect to the prior distribution over the set of models. The higher the KL scores, the more discriminate the detection measures are. Surprises identified by the computational model turned out to have a good match to human viewers' eye movements on video images.

Later in this chapter, we will discuss an example of identifying novel as well as salient themes in the literature of terrorism research so that it helps analysts to identify not only high-profile topics but also low-profile topics overshadowed.

Surprises are surprises because they are not expected in a particular context. Similarly, there is an interesting connection between context and creativity. An idea could be seen as trivial or common in one community but inspirationally creative in another. Where do we expect to find creative ideas in our society, an environment full of information patches? This question has been addressed from a social capital point of view based on a concept called *structural holes*.

5.1.4 Structural Holes and Brokerage

Structural holes are defined as a topological property of a social network. According to Burt (2004), structural holes refer to the lack of comprehensive connectivity among components in a social network. The connection between distinct components is reduced to few person-to-person links at structural holes, whereas person-to-person links tend to be uniformly strong at the center of a group. Because information flows are restricted to the privileged few who are strategically positioned over structural holes, the presence of a structural hole has a potential for gaining distinct advantages.

Opportunities generated by structural holes are a vision advantage. People connected across groups are more familiar with alternative ways of thinking, which gives them more options to select from and synthesize. Because they have more alternative ideas to choose from, the quality of the selection tends to be better.

Burt identifies four levels of brokerage through which one can create value, from the simplest to the most advanced:

1) Increasing the mutual awareness of interests and difficulties of people on both sides of a structural hole.
2) Transferring best practice between two groups.
3) Drawing analogies between groups seemingly irrelevant to one another.
4) Synthesizing thinking and practices from both groups.

Burt found indeed that a vision advantage associated with brokerage translates to better received ideas.

A brokerage between information foraging and the structural holes theory may be fruitful for visual analytics. The structural-hole induced brokerage perspective addresses situations where information scents are either missing or unreachable. On the other hand, the structural-hole theory can guide the selection of potentially information-rich paths for foragers. Consider the three prominent clusters shown in Fig. 5.1, a brokerage-oriented perspective focuses on the linkages connecting distinct clusters. Consequently, the focus on inter-cluster connections provides us a unique leverage to differentiate individual papers at a higher level of aggregation. In the terrorism research example, the earliest theme is about physical injuries; a later theme is centered on health care and emergency responses; the most recent theme focuses on psychological and psychiatric disorders. Cognitive transitions from one theme to another become easier to grasp at this level. This high-level understanding can also serve as a meaningful context for us to detect what is common and what is surprising. To understand terrorism research as a whole, it is necessary to understand how these themes are interrelated. The whole here is indeed more than the sum of parts. An information-theoretic view brings us a macroscopic level of insights.

5.1.5 Macroscopic Views of Information Contents

The following example illustrates an information-theoretic approach to the analysis of low-profile thematic patterns as well as high-profile thematic patterns.

Information entropy is a useful system-level metric of fluctuations of the overall information uncertainty in a large-scale dynamic system. Fig. 5.2 shows how the information entropy of terrorism research changes over 18 years based on keywords assigned to scientific papers on the subject between 1990 and the first half of 2007. Entropies are computed retrospectively based on the accumulated vocabulary throughout the entire period. Two consecutive and steep increases of entropy are prominently revealed, corresponding to 1995 – 1997 and 2001 – 2003. The eminent increases of uncertainty send a strong message that the overall landscape of terrorism research must have been fundamentally altered. The unique advantage of the information-theoretic insight is that it identifies emergent macroscopic properties without overwhelming analysts with a large amount of microscopic details. Using the terminology of information foraging, these two periods have transmitted the strongest information scent. Note that using the numbers of unique keywords fails to detect the first period identified by information entropy. Subsequent analysis at microscopic levels reveals that the two periods are associated with the Oklahoma City bombing in 1995 and the 911 terrorist attacks in 2001.

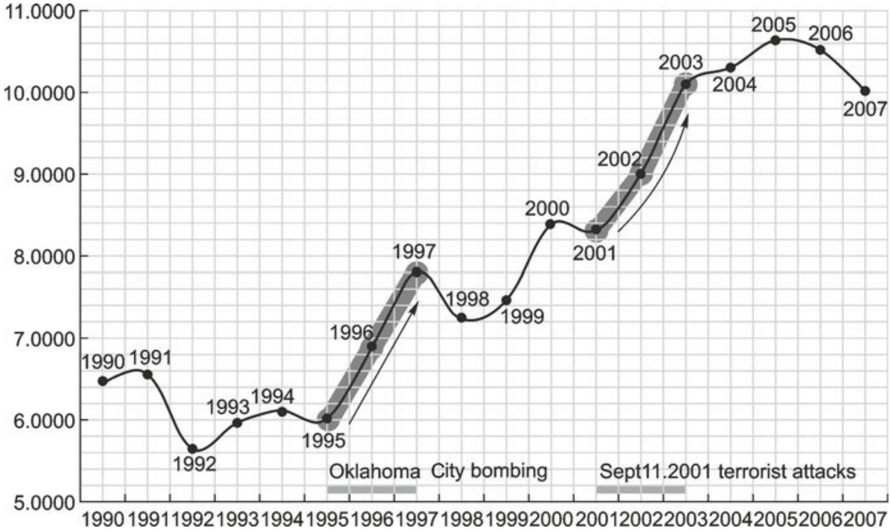


Fig. 5.2 Information entropies of the literature of terrorism research between 1990 and the first half of 2007. The two steep increases correspond to the Oklahoma City bombing in 1995 and the 911 terrorist attacks in 2001. Source: (Chen, 2008).

Information indices allow us to compare the similarity between different years. Fig. 5.3 shows a 3D surface of the K-L divergences between distributions in different years. The higher the elevation is, the more difference there is between two years of research. For example, the blue area has the lowest elevation, which means that research is more similar in the recent three years than earlier years.

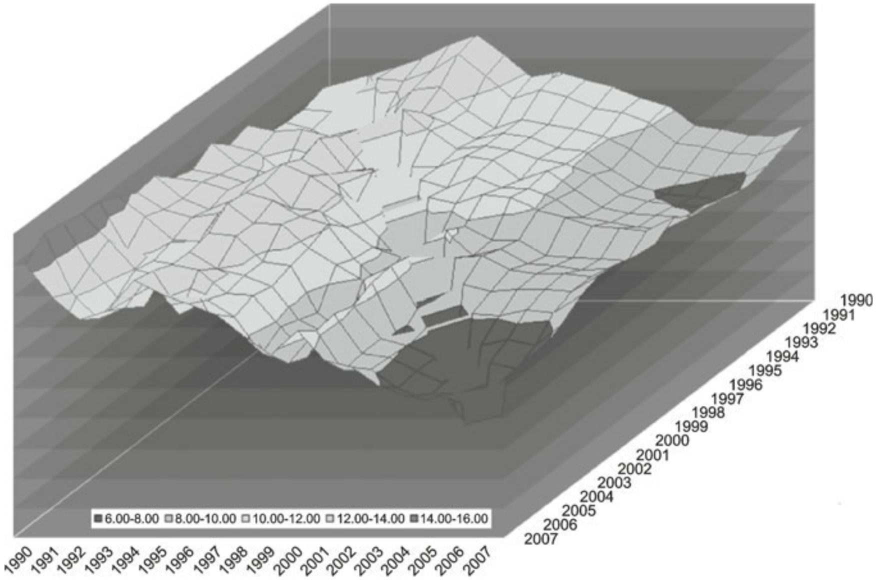


Fig. 5.3 Symmetric relative entropy matrix shows the divergence between the overall use of terms across different years. The recent few years are most similar to each other. The boundaries between areas in different colors indicate significant changes of underlying topics. Source: (Chen, 2008). (see color figure at the end of this book)

Information-theoretic techniques provide not only a means of addressing macroscopic questions, but also a way to decompose and analyze questions at lower levels of aggregation. Given that we have learned that there are two distinct periods of fundamental transformation in terrorism research, the next step is to understand what these changes are in terms of their saliency and novelty. Different distributions may lead to the same level of entropy. In order to compare and differentiate different distributions, one can use information-theoretic metrics such as information bias, which measures the degree to which a sub-sample differs from the entire sample that the sub-sample belongs to. High-profile thematic patterns can be easily identified in terms of term frequencies. Low-profile thematic patterns are information-theoretic outliers from the mainstream keyword distributions. Low-profile patterns are equally important as high-profile patterns in analytical reasoning because they tell us something that we are not familiar with, and something novel.

Informational bias $T(a{:}B)$ is defined as follows, where a is a sub-sample of the entire sample B. p_{ab}, p_a, p_b, and $p_{b|a}$ are corresponding probabilities and conditional probabilities. $H_a(B)$ is the conditional entropy of B in the sub-sample. We take the distribution of a given keyword and compare it to the entire space of keyword distributions.

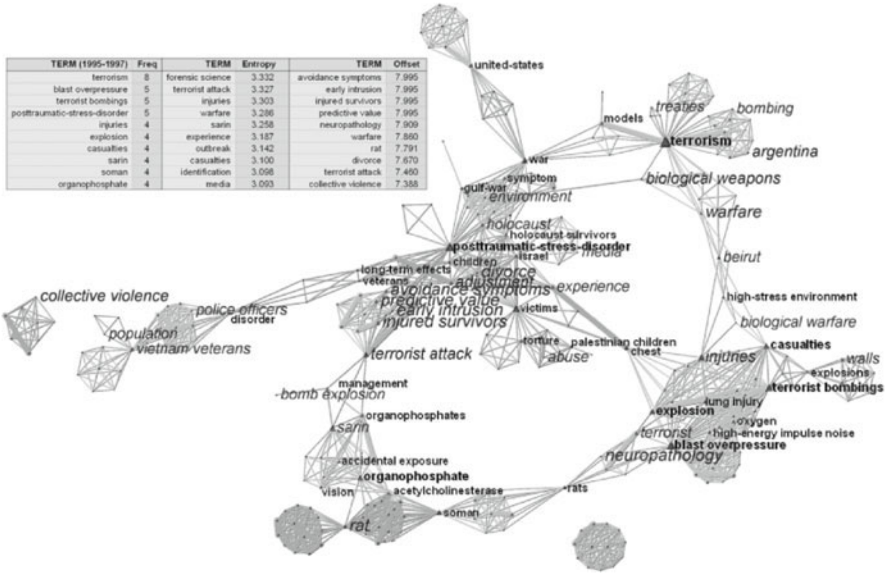
$$T(a:B) = \frac{1}{p_a}\sum_b p_{ab}\log_2\frac{p_{ab}}{p_a p_b}$$

$$T(a:B) = \sum_b p_{b|a}\log_2\frac{p_{b|a}}{p_b} = -\sum_b p_{b|a}\log_2 p_b - H_a(B)$$

Fig. 5.4 illustrates how one may facilitate a sensemaking process with both high- and low-profile patterns embedded in the same visualization. The network in Fig. 5.4 consists of keywords that appeared in 1995, 1996, and 1997, corresponding to the first period of substantial change in terrorism research. High-profile patterns are labeled in black, whereas low-profile patterns are labeled in dark red. High-profile patterns help us to understand the most salient topics in terrorism research in this period of time. For example, terrorism, posttraumatic-stress-disorder, terrorist bombings, and blast overpressure are the most salient ones. The latter two are closely related to the Oklahoma city bombing event, whereas *posttraumatic-stress-disorder* is not directly connected at this level. In contrast, low-profile patterns include *avoidance symptoms*, *early intrusion*, and *neuropathology*. These terms are unique with reference to other keywords. Once these patterns are identified, analysts can investigate even further and make informed decisions. For example, one may examine whether this is the first appearance of an unexpected topic or whether the emergence of a new layer of uncertainty to the system at this point makes perfect sense.

Developing methods and principles for representing data quality, reliability, and certainty measures throughout the data transformation and analysis process is a key element on the research agenda for visual analytics (Thomas & Cook, 2005). Each of the individual method illustrated here has been used in their own application domains and some of them have already been applied to visual analytics. However, introducing the collection of theories, strategies, and techniques as a consistent and yet versatile information-theoretic view of visual analytics is expected to strengthen the theory and practice of visual analytics.

At the beginning of the chapter, we emphasize that asking the right question holds the key to connecting the dots. Examples discussed here illustrate various ways to find the dots, make sense of the dots, and differentiate dots at different levels of abstraction, ranging from macroscopic to microscopic levels. The information-theoretic perspective provides a potentially effective framework to address questions concerning analytical reasoning with uncertainty, synthesizing evidence from multiple sources, and developing a macroscopic understanding of a complex, large-scale, and diverse body of information



TERM (1995-1997)	Freq	TERM	Entropy	TERM	Offset
terrorism	8	forensic science	3.332	avoidance symptoms	7.995
blast overpressure	5	terrorist attack	3.327	early intrusion	7.995
terrorist bombings	5	injuries	3.303	injured survivors	7.995
posttraumatic-stress-disorder	5	warfare	3.286	predictive value	7.995
injuries	4	sarin	3.258	neuropathology	7.909
explosion	4	experience	3.187	warfare	7.860
casualties	4	outbreak	3.142	rat	7.791
sarin	4	casualties	3.100	divorce	7.670
soman	4	identification	3.098	terrorist attack	7.460
organophosphate	4	media	3.093	collective violence	7.388

Fig. 5.4 A network of keywords in the terrorism research literature (1995 – 1997). High-frequency terms are shown in black, whereas outlier terms identified by informational bias are shown in dark red. Source: (Chen, 2008). (see color figure at the end of this book)

systematically. The information-theoretic perspective is expected to stimulate further advances of visual analytics and work harmoniously with other approaches to facilitate analytical reasoning.

5.2 Turning Points

The late sociologist Murray S. Davis developed some intriguing insights into why we are interested in some (sociological) theories but not others (Davis, 1971a). Although his work focused on sociological theories, the insights are broad-ranging. According to Davis, "the truth of a theory has very little to do with its impact, for a theory can continue to be found interesting even though its truth is disputed — even refuted!" A theory is interesting to the audience because it denies some of their assumptions or beliefs to an extent. But if a theory goes beyond certain points, it may have gone too far and the audience will lose their interest. People pay attention to a theory not really because the theory is true and valid; instead, a theory is getting people's attention because it may change people's beliefs.

5.2.1 The Index of the Interesting

Something interesting is what engages our attention. Davis masterfully prompted us with the question: where was the attention before it was engaged by the interest? Before the state of attention, most people are not attentive to anything in particular and we take many things for granted. Harold Garfinkel (1967) called this state of low attention "the routinized taken-for-granted world of everyday life." If some audience finds something interesting, it must have stood out in their attention in contrast to the taken-for-granted world — it stands out in their attention in contrast to propositions that a group of people have taken for granted. The bottom line is that a new theory will be noticed only when it denies something that people take for granted, such as a proposition, an assumption, or a commonsense.

One of the keynote speakers at IEEE VisWeek 2009 at Atlantic city told the audience how to tell an engaging story. The template he gave goes like this: here comes the hero; the hero wants something, but the hero is prevented from getting what he wants; the hero wants it badly; and finally, the hero figures out an unusual solution. The storytelling rubric has something in common with why a theory is interesting. Both are connected to our curiosity of something unexpected.

Davis specifies a rhetorical routine to describe an interesting theory: 1) The author summarizes the taken-for-granted assumptions of his audience. 2) The author challenges one or more conventional propositions. 3) The author presents a systematically prepared case to prove that the wisdom of the audience is wrong. He/she also presents new and better propositions to replace the old ones. 4) Finally, the author suggests the practical consequences of new propositions.

David realized the nature of interesting can be explained in terms of a dialectical relation between what appears to be and what could really exist. In his terminology, the appearances of a phenomenon experienced through our senses are phenomenological, whereas intrinsic characteristics of the phenomenon are ontological. An interesting theory, therefore, is to convince the audience that the real nature of a phenomenon is somewhat different from what it seems to be. He identified 12 logical categories that capture such dialectical relations, shown in Table 5.1.

Table 5.1 Dialectical relations between taken for granted assumptions and new propositions.

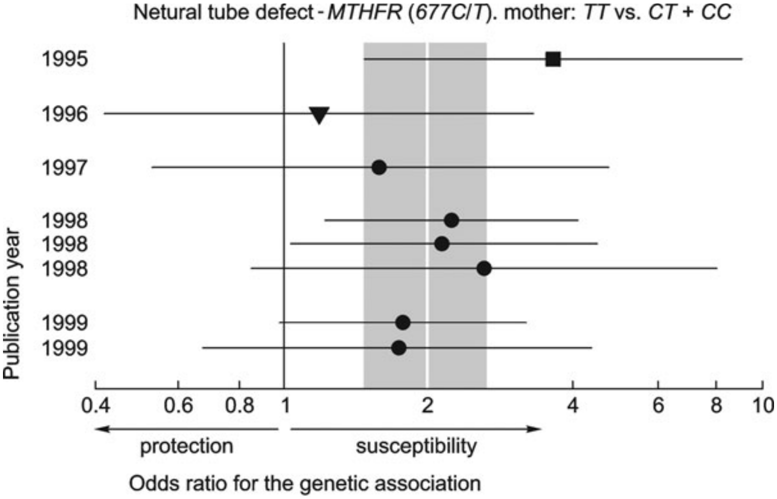
Phenomenon		Dialectical Relations		
Single	Organization	Structured	⟵⟶	Unstructured
	Composition	Atomic	⟵⟶	Composite
	Abstraction	Individual	⟵⟶	Holistic
	Generalization	Local	⟵⟶	General
	Stabilization	Stable	⟵⟶	Unstable
	Function	Effective	⟵⟶	Ineffective

Continued

Phenomenon		Dialectical Relations		
	Evaluation	Good	⟷	Bad
Multiple	Co-relation	Interdependent	⟷	Independent
	Co-existence	Co-exist	⟷	Not co-exit
	Co-variation	Positive	⟷	Negative
	Opposition	Similar	⟷	Opposite
	Causation	Independent	⟷	Dependent

5.2.2 Proteus Phenomenon

Proteus is a sea god in Greek Mythology. He could change his shape at will. The Proteus phenomenon refers to early extreme contradictory estimates in published research. Controversial results can be attractive to investigators and editors. Ioannidis and Trikalinos (2005) tested an interesting hypothesis that the most extreme, opposite results would appear very early as data accumulate rather than late. They used meta-analyses of studies on genetic associations from MEDLINE and meta-analyses of randomized trials of health care interventions from the Cochrane Library. They evaluated how the between-study variance for studies on the same question changed over time and at what point the studies with the most extreme results had been published. The results show that for genetic association studies, the maximum between-study variance was more likely to be found early in the 44 meta-analyses and 37 in the health care interventions case. The between-study variance



(a)

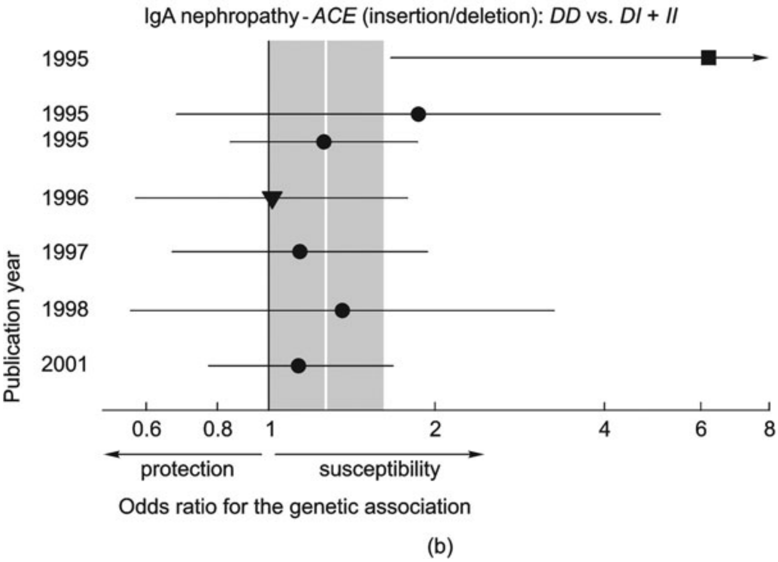


Fig. 5.5 The swing of results decreased over time. Source: Figure 1 in (Ioannidis & Trikalinos, 2005).

decreased over time in the genetic association studies, which was statistically significant (Fig. 5.5). This 2005 study itself has attracted 330 citations in 2010.

5.2.3 The Concept of Scientific Change

The nature of scientific change has been studied in the philosophy of science (Collins, 1998; Laudan, Donovan, Laudan, Barker, Brown, Leplin, Thagard, & Wykstra, 1986; Schaffner, 1992), sociology (Fuchs, 1993; Griffith & Mullins, 1977), and history of science (Brannigan & Wanner, 1983). Quantitative studies of the topic can be found in the fields of scientometrics, citation analysis, and information science in general (Chen, 2003; Heinze & Bauer, 2007; Heinze, Shapira, Senker, & Kuhlmann, 2007; Hummon & Doreian, 1989; Small & Crane, 1979; Sullivan, Koester, White, & Kern, 1980; Wagner-Dobler, 1999). Scientific literature has increasingly become one of the most essential sources for these studies. Social network analysis and complex network analysis also provides valuable perspective (Barabási, Jeong, Néda, Ravasz, Schubert, & Vicsek, 2002; Newman, 2001; Redner, 2004; Snijders, 2001; Valente, 1996; Wasserman & Faust, 1994).

It is evident that scientific discoveries share important and generic properties (Bradshaw, Langley, & Simon, 1983; Simon, Langley, & Bradshaw, 1981a). In order to obtain conclusive evidence, one will need a theory of scientific discovery that can provide a unifying conceptual framework so that one can characterize a variety of scientific discoveries from a consistent perspective. In addition, given one concrete case of scientific discovery, it may be studied from multiple and often not interconnected perspectives. For example, a philosophical study of a scientific revolution may have little overlap with a sociological study of the same process. Even two philosophical studies of the same scientific revolution could appear to be unrelated in the eyes of laypersons. Statistical models of network evolution have been used to identify statistical and topological properties of scientific networks. However, such properties, although generic in nature, do not readily offer further explanations of why scientists in a network behave in a particular way. Motivations, decisions, and interpretations underlying such properties are often detached or left out. Thus, we need a theory that not only identifies statistical and topological properties of scientific networks, but also offers practical insights into the mechanisms that may drive scientists' observed behavioral patterns.

There are many types of theories, including descriptive, explanatory, generative, predictive, and prescriptive (Bederson & Shneiderman, 2003). A simple, descriptive, explanatory, and generative theory of scientific discovery is ideal based on generic mechanisms of discovery. Such generic mechanisms are in fact generative in nature because scientists and computer simulation algorithms would be able to emulate such mechanisms. We have a few expectations of our new theory. First, it should help us to recognize the significance of new discoveries as soon as possible. Second, it should help us to identify as many potential areas of growth as possible. Third, it should help us to explain both the creation of knowledge and its diffusion within a consistent and unified framework.

A review of the literature on scientific change in the philosophy of science, sociological theories of scientific change, sociological theories of creative ideas, information foraging theory converged to a recurring theme. The recurring theme is that insights, creative ideas, and transformative scientific discoveries are the work of a broad range of brokerage and boundary spanning mechanisms.

Building on this recurring theme, we construct a simple theory of scientific discovery to explain why transformative research is likely to be created by this type of brokerage mechanisms. One can derive many interesting conjectures from the first principles of the theory, including structural and temporal properties of citation and co-citation networks. In particular, we will show that the theory can reduce a large number of variables studied in the literature with the boundary spanning paradigm.

5.2.4 Specialties and Scientific Change

Specialty is a key concept in the study of scientific change. A specialty is a group of researchers and practitioners who have similar training, attend the same conferences, read and cite the same body of literature (Fuchs, 1993). There are a variety of studies of specialty in the literature (Chubin, 1976; Fuchs, 1993; Morris & Van der Veer Martens, 2008; Mullins, Hargens, Hecht, & Kick, 1977; Small & Crane, 1979). For example, Mullins et al. studied author groups corresponding to co-citation clusters using questionnaires and concluded that co-citation clusters indeed represent the intellectual structure and that coauthors do form social groups (Mullins et al., 1977). Co-author networks have also been studied in complex network analysis of community structures (Girvan & Newman, 2002). These finding provide an empirical basis for the analysis of scientific change based on co-citation networks.

The dynamics of the structure of a specialty is a central issue in the context of scientific change. Research has shown that major changes in a variety of disciplines tend to be originated within small, socially coherent groups (Griffith & Mullins, 1977). Kuhn observed that new paradigms are typically initiated by young scientists or newcomers to a crisis-laden field (Kuhn, 1962). In addition, Crane (1969) found that the desire for originality motivates scientists to maintain contacts with scientists and scientific work in areas different from their own in order to enhance their ability to develop new ideas in their own areas. This observation underlines an intriguing fact that many major scientific discoveries are often fundamentally inspired by external influences, or from peripheral areas of established research specialties, which echoes Kuhn's earlier observation.

Crane's observation can be seen as a special case of what sociologist Burt called the social capital of *structural holes* (Burt, 1992, 2001, 2004). Structural holes are voids in social structure. According to Burt's theory of structural hole, structural holes in a social network are disconnected or poorly connected areas between tightly and densely connected groups of people. The presence of such structural holes may influence the importance of positions in a social network — some positions become more privileged and competitive than others. The value of a person in a social network is therefore linked to the potential to establish connections between groups that are separated by structural holes. People in positions with great brokerage potentials are known as brokers and gatekeepers. Brokers are rewarded for their integrative work in terms of more positive evaluations, higher compensations, and faster promotion. The underlying reason for the difference is that information is more homogeneous within groups, whereas more heterogeneous between groups. Brokers are in special positions to access heterogeneous information from a broader range of sources. The privilege often leads to competitive advantage. In the following sections, we will argue that the role of brokerage mechanisms not only goes beyond social networks, but also underlines an important source of insight that leads to profound scientific changes and discoveries.

The dynamics of the theory-change in science is not only a philosophical issue, but also a historical one. Brush investigated whether scientists give greater weight to novel predictions than to explanations of known facts against historical cases in physical science (Brush, 1994). Several theories were accepted after successful novel predictions but there is little evidence that extra credit was given for novelty. Others were accepted without making successful novel predictions. No examples were found of theories that were accepted primarily because of successful and novel predictions and would not have been accepted if those facts had been previously known. Brush further examined the impact of predictions on theory acceptance through several cases, including the Big Bang vs. steady-state cosmology, the origin of the Moon, gravitational light bending, and Hannes Alven's plasma physics (Brush, 1995). Brush concluded that confirmed predictions provide "corroboration" of a hypothesis, but only in the minimalist sense of scientists voting with their publications. Corroboration "merely makes it more reasonable to pursue that hypothesis than one that has not been corroborated," and thus "there was a significant increase in publications on the theory [i.e., those theories in the case studies] that led to the prediction".

A mathematical approach to the prediction of scientific discovery was proposed in (Goffman & Harmon, 1971). Their approach is built on a four-state Markov chain model of discovery. Discovery is conceptualized as a process of placing a set of information in the right order. They were able to construct such a model based on an expert-annotated bibliography of the field of symbolic logic. The discovery per se would be the ordered information, i.e. patterns. The four states are defined in terms of the sufficiency and order of information. In State I, information is insufficient and unordered. The problem at this stage is to acquire information, not to order it. Observations are inadequate to establish patterns. In State II, information is insufficient but available information is ordered. In State III, there are sufficient information elements, but not in the right order. Finally, in state IV, information is both sufficient and ordered. The discovery is established. From here, it can be elaborated, refined, or challenged.

5.2.5 Knowledge Diffusion

Mark Twain thought he could learn how to become a Mississippi river pilot by studying charts and manuals. In fact, he discovered that it would take several years of apprenticeship to become an experienced river pilot and countless of journeys over the same terrain to be able to "read" the meaning of currents and water-levels of the ever-changing river in always-different circumstances (Twain, 2001).

Diffusion depends on three things:

1) Is the information interesting enough to share with others?
2) How much will it cost for the messenger to send the information across?
3) How easy can it be passed from one to another?

Morten Hansen's work on the role of weak ties in knowledge sharing is a boundary spanning effort in itself (Hansen, 1999). Hansen combines the concept of weak ties from social network research and the notion of complex knowledge in organizational knowledge management to explain the role of weak ties in sharing knowledge across the boundaries of organizational units. The study found that neither weak nor strong relationships between operating units lead to efficient sharing of knowledge among them. When the knowledge is highly complex, strong ties provide the highest effect, whereas weak ties have the strongest effect when the knowledge is not complex.

Research in several fields distinguishes two forms of knowledge: explicit and tacit. The most typical example of explicit knowledge is science, whereas the most common tacit knowledge is art. The main issue here is that it seems much harder to transfer tacit knowledge than transferring explicit knowledge. The notion of tacit knowledge is originally developed by Michael Polanyi (1958, 1966). Tacit knowledge is hard to articulate. In many situations, it can only be acquired through experience. An informal way to characterize tacit knowledge is "we know more than we can tell". For example, we can recognize people by their face as a whole, but if we focus on specific features of a face, the face recognition is taken over by something quite different. Similarly, if a dancer focuses on the details of the elements of the performance, the performance itself would fall apart.

The inventors of the jet engine, hovercraft, and numerous others have found out that the world was skeptical and reluctant to be convinced. Many have asked the same question: why and what distinguished these inventions from others that were promptly accepted and brought into successful production. Little attention has been paid to the detailed scientific history of individual technological innovations, especially the specific circumstances and steps which led to the acceptance of a new product or a process. Cahn (1970) described why case histories of technological inventions as well as scientific discoveries were valuable for people to find their way through the complexities of modern science and technology. In particular, a useful kind of case histories should cast light on the mechanism of connecting scientists and technologies.

There is a growing interest recently in how information spreads over web logs, or blogs (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004). Burst detection was applied to the discovery of sub-structures in blogspace based on spiking inter-blog links (Adar, Zhang, Adamic, & Lukose, 2004; Gruhl et al., 2004; Kumar, Novak, Raghavan, & Tomkins, 2003, 2004).

A widely known model of information diffusion in society is the *two-step flow model*, which holds that the spread of information from mass media to the society takes two steps. First, the information is filtered through opinion leaders, who then influence others (Katz & Lazarsfeld, 1955; Lazarsfeld, Berelson, & Gaudet, 1944). It is important to understand the role of opinion leaders in scientific communities. Co-authorship, apprenticeship, and geographic proximity are among the strongest types of ties for knowledge diffusion.

Plate tectonics in geology is a recent example of a scientific revolution (Stewart, 1990; Thagard, 1992). Two key elements of plate tectonics are continental drift and seafloor spreading theories. There are well-documented sociological studies of the acceptance of continental drift and seafloor spreading theories, especially the use of citation analysis in a sociological context (Stewart, 1990).

Continental drift was first proposed as early as 1912 by Alfred Wegener in attempts to explain the apparent matching coastline shapes between Africa and South America, For decades geologists did not accept his theory because given the knowledge of geology at that time it was not conceivable how continental drift could take place. The discovery of the Mid-Atlantic Ridge in the 1950s was a pivotal event that led to the theory of seafloor spreading and general acceptance of Wegener's theory of continental drift. The theory of continental drift was not widely accepted in Europe until the 1950s, but it was not accepted by geologists in North America until 1960s.

One of the reasons that Wagner's continental drift theory was initially rejected was because he didn't have a convincing explanation of why and how the continents move.

Quantitative models of how scientific ideas spread are proposed by many researchers (Bettencourt, Castillo-Chavez, Kaiser, & Wojick, 2006; Bettencourt, Kaiser, Kaur, Castillo-Chavez, & Wojick, 2008). Epidemic models are among the most popular ones (Goffman & Newill, 1964; Liben-Nowell & Kleinberg, 2008; Nowakowska, 1973). Epidemic models consider variables such as contact rates between scientists, latency and recovery times. The contact rate between scientists is found to be the single important factor to speed up the diffusion of knowledge.

Other potentially applicable models of diffusion include ant colony and random walk models. In an ant colony model (Dorigo & Gambardella, 1997), ants travel between their home and food sources. They leave scents as trails for others. Scents decrease over time unless being reinforced by other ants. One can see a natural mapping from the ant colony model to a model of network evolution. Here ants are replaced by scientists. Their home is now the contemporary intellectual structure. The food sources are new publications in the literature. Finding foods is equivalent to making a reference to a new publication. Doing so also leaves trails for other scientists. Ant colony is a self-organizing optimization mechanism. Unlike the preferential attachment approach, which specifies the criteria of an addition to an existing network, ant colony is not limited to preferences, although it can be tailored to make use of them.

Random walk algorithms are also useful for modeling the spread of information. A random walk model over a network is built on state transition probabilities. Each node in the network represents a state. Moving from one node to another is governed by a state transition probability, which can be updated based on available evidence in Bayesian rules. The spread of knowledge is thus translated into a question of how easy or how hard it would be to make such moves.

The ant colony and random walk models have a more profound connection to the *information foraging theory* (Pirolli, 2007). The fundamental premise of the information foraging theory is that the behavior of a forager, namely, information searchers and, in this case, scientists is driven by a perceived or calculated profitability of the potential move. The profitability takes into account the expected returns as well as potential risks or costs involved. For example, if online access to an article costs $30, then the cost is only part of the equation. Whether the article is worth your paying the $30 depends on what you can do with the article and how urgently you need it.

Sandstrom argued that information seekers are very much like foragers, especially in terms of how and where they forage for valued resources (Sandstrom, 1999). She introduced the notion of bibliographic microhabitats to underline the similarity between hunters and information seekers. She further argued that if some empirical cost-benefit currency can be established, then analysts would be able to rank foragers' preferences, predict which resources will be pursued, and specify the net returns associated with particular choices.

In summary, unlike epidemic models, foraging models emphasize not only structural properties of an information space for information seekers or a problem space for scientists, but also the interplay between perceived values, handling costs, and various competing and probably conflicting factors in a broader context of decision making. In other words, one may incorporate foraging models into existing workflows so that one can recognize and act upon vital clues that may lead them to a fruitful path.

5.2.6 Predictors of Future Citations

Whenever we do not have sufficient expertise to make our own judgment, we resort to the opinions of experts, track records of the past, referrals and recommendations from friends, or reputations and brands. One lesson learned from the cases of September-11 terrorist attacks and Iraqi's WMD is that if we draw our conclusions from indirect evidence and inference, then it is crucial to make it clear. It is a valuable lesson because it is also applicable to many situations in which first-hand evidence is not available and we have to rely on indirect measurements. If we know little about the quality and credibility of a scientific paper, what are the factors that will influence whether or not the paper will be cited?

Predicting future citations is a topic of interest for researchers, evaluators, and science policy makers. The predictive power of a diverse range of variables has been tested in the literature. As shown in Table 5.2, most of the commonly studied variables can be categorized into a few groups according to their parent classes where they belong to. For example, the number of pages of a paper is an attribute of the paper as an article. The number of authors of a paper is an attribute of the authorship of the paper. One can expect even more variables will be added to the list. We expect to demonstrate that our theory of transformative discovery provides a theoretical framework to accommodate this diverse set of attributive variables and provides a consistent explanation for most of them.

Table 5.2 Variables associated with articles that may be predictive of their subsequent citations.

Components	Attributive Variables	Hypotheses derived from theory of discovery
Article	Number of pages	Boundary spanning needs more text to describe.
	Number of years since publication	
Authorship	Number of authors	More authors are more likely to contribute from diverse perspectives.
	Reputation (citations, h-index)	
	Gender	
	Age	
	Position of last name in alphabet	
Impact	Citation counts	The value of the work is recognized.
Usage	Download times	The value of the work is recognized.
Abstract	Number of words	Transformative ideas tend to be more complex than simple ones. More words are needed to express more complex ideas.
	Structured (yes/no)	
Content	Type of contributions: tools, reviews, methods, data, etc.	
	Scientific rigorous of study design	
Reference	Number of references	More references are needed to cover multiple topics that are being synthesized.
Discipline	Number of disciplines	It is more likely that the work synthesizes multiple disciplines.
Country	Number of countries	It is more likely that authors from different countries bring in distinct perspectives.

Continued

Components	Attributive Variables	Hypotheses derived from theory of discovery
Institution	Number of institutions	It is more likely that authors from different institutions bring in distinct perspectives.
Journal	Impact factor	
	Indexed by different databases	
Sponsored	yes/no	

Several studies focus on the relationship between earlier download times and subsequent citations (Brody, Harnad, & Carr, 2006; Lokker, McKibbon, McKinlay, Wilczynski, & Haynes, 2008; Perneger, 2004). Perneger (2004) studied 153 papers published in one volume of the journal BMJ in 1999 (volume 318) along with the full paper download times within the first week of publication and their citations as of May 2004 recorded in the Web of Science. Perneger coded each paper in terms of its study design using 7 categories, namely randomized trials, systematic reviews, prospective studies, case-control studies, cross sectional surveys, qualitative studies, and other designs. He found a statistically significant positive Pearson correlation of 0.50 ($p<0.001$) between citations and the download times within the first week. He also found that 33% of variance can be explained by hits (download times) and the length of a paper. A correlation of 0.4 was found between citations and downloads of articles in the e-print archive repository arXiv (Brody et al., 2006), although the amount of variance explained (16%) was relatively low.

In a more recent analysis, a group of researchers at McMaster University, Canada, studied whether 20 article and journal variables can predict citations of 1,274 articles from 105 journals published between January and June 2005 (Lokker et al., 2008). The 20 variables include ratings of clinical relevance and newsworthiness, which are routinely collected by the McMaster online rating of evidence system. The dataset was split by 60:40 for derivation and validation. Their study shows that a multiple regression model accounted for 60% of the variance in the derivation portion of the dataset. The same model accounted for 56% of the variance in the validation dataset. Higher citations were predicted by indexing in numerous databases, number of authors, number of cited references, clinical relevance scores, original papers, multi-center studies, and a few other variables.

Dalen and Kenkens (2005) studied 1,371 articles published in 1990 – 1992 in 17 demography journals in order to identify explanatory factors that may influence the visibility of an article. In particular, they were interested in whether the reputations of authors and journals had anything to do with the citations these papers received later on and how soon they would get their first citation. An author's reputation was estimated based on the citations of the author in 1990, the first year of the period. If an article has multiple authors, the most prominent author's reputation was used. The reputation of a journal was represented by its impact factor in 1990. They used duration analysis, originated in survival analysis, to analyze the data. The central question is: what determines the probability of an article changing from the initial state of not being cited to a state in which it is cited? In survival analysis, the role of a hazard function is to estimate the probability of transitions from the initial state. The simplest form of a hazard function is constant with no memory of how long the initial state lasts. In other words, the probability of an article moving away from the initial state in the next time frame does not depend on how much time it has been spent in the initial state. More realistic hazard functions include positive and negative duration dependence. Positive duration dependence specifies that the longer an article has been in the initial state, i.e. not being cited, the better the chance it will be cited. In contrast, negative duration dependence assumes the opposite. Dalen and Henkens chose their hazard function based on the Gompertz distribution.

The Gompertz distribution is a theoretical distribution of survival times. It was proposed by Gompertz in 1825 to model human mortality. The resultant hazard function is defined as follows:

$$y(t) = ae^{be^{ct}}$$

where a is the upper asymptote, i.e. the value of $y(\mathrm{t} \to \infty)$ in the infinite future time, b is the x displacement, c is the growth rate, and e is the Euler's number. The Gompertz function models the slow growth at the initial and final stages and faster growth in intermediate stages. It has been used to model the growth of tumors, the uptake of mobile phones, and the mortality of population.

Dalen and Henkens divided articles into four categories and then used a statistical method called multinomial logit to test how explanatory factors such as authors' and journals' reputations could explain the citation patterns.

1) Articles with citations too little and/or too late (forgotten ones).
2) Articles with late citations (sleeping beauties).
3) Articles with early citations but fading off quickly (flash-in-the-pans).
4) Articles with early citation and many subsequent cites (normal science).

$$\text{Prob (Article = sleeping beauty)} = \exp(X\beta^{(2)}) / \\ [1 + \exp(X\beta^{(2)}) + \exp(X\beta^{(3)}) + \exp(X\beta^{(4)})]$$

Their model shows statistically significant effects of several explanatory variables such as author reputation, the number of pages, and journal reputation (impact factor) at $p < 0.01$.

The survival model of the timing of first citation identified the major role of the communication process in speeding up the uptake of a scientific paper, namely visibility, language and reputation of authors and journals. When the effect of a journal's quality such as the reputation of the editors and editorial policy is controlled, the duration analysis reveals the reputation effect of authors. The effect of journals becomes clear.

Dalen and Henkens' duration study essentially tell us that the reputation of the authors of an article and the reputation of the journal in which the article is published are the most critical factors for the article to gain visibility and get cited. Are we attracted by other signals? What about structural, temporal, and semantic properties of the underlying topic?

What is the extent to which quantitative rankings of highly cited authors confirm or, even more ambitiously, predict Nobel Prize awards? Between 1977 and 1992, Eugene Garfield published a series of studies of Nobel Prize winners' publications and their citations and made predictions of future Nobel Prize laureates based on existing citation data.

He reported that eight Nobel laureates were found on a list of 100 most cited authors from 1981 through 1990(Garfield & Welljamsdorof, 1992). Others on the list were seen as potential Nobel Prize winners in the future. On the other hand, it was noted that the undifferentiated rankings of the most cited authors in a given period of time could be further fine-tuned to increase the accuracy of its coverage of Nobel Prize awards. For example, the Nobel Committee sometimes selects relatively small specialties. Further dividing the list according to specialties shows that Nobel laureates in relatively small specialties are among the most cited authors in their specialties.

Methods papers of Nobel Prize winners tend to attract a disproportionably high amount of citations. More recent examples of methodological contributions include the 2007 Nobel Prize for the British embryonic stem cell research architect Martin Evans. Garfield coined the phenomenon the *Lowry Phenomenon*, referring to the classic example of Oliver Lowry's 1951 methods paper, which was cited 205,000 times up to 1990.

Research has shown that citation frequency has a low predictive power for Nobel awards because there are so many other scientists with the same or even higher citations as the few Nobel Prize winners. The greatest value of counting citations is its simplicity. Subsequent attempts to improve the accuracy of the method tend to lose the simplicity. Hirsch's h-index has drawn much interest also because of its simplicity despite its known limitations (Hirsch, 2005a). Antonakis and Lalive intended to capture both the quality and productivity of a scholar with a new index IQp (Antonakis & Lalive, 2008). They compared the new index of Nobel winners in physics, chemistry, medicine, and economics. It is worth noting here that one should always be cautious when using quantitative indicators in qualitative decisions. The authors found about two third of Nobel winners have an IQp over 60. The authors showed that in several examples, IQp differentiated Nobel class and others more accurately than the h-index, including physicist Ed Witten (h=115 and IQp=230) and others who have high h-index but relatively low IQp index, S. H. Snyder (h=198, IQp=117) and R. C. Gallo (h=155, IQp=75).

In the context of scientific discovery, we will expand the information foraging theory to describe the behavior of scientists in searching for novel hypotheses and theories. This will help us to explain how a scientist would maximize the profitability of the next move.

5.3 Generic Mechanisms for Scientific Discovery

There is evidence in the literature that scientific discoveries do share some common mechanisms, especially in light of research in computer simulation of discoveries, cognitive studies of scientific change, and the nature of insight.

5.3.1 Scientific Discovery as Problem Solving

The most prominent work in this area has been done by Herbert Simon and his colleagues using computer simulation to study and reconstruct scientific discoveries (Bradshaw et al., 1983). A long list of examples of automated discoveries was given in (Glymour, 2004). He used the metaphor of finding a needle in a haystack to characterize the problems faced by scientists in discovery. Rather than sifting through things in the haystack one by one, automated discovery is akin to either setting the haystack on fire and blowing away the ashes to find the needle, or running a magnet through the haystack. There are advantages and limitations. Following the metaphor, for example, the fire may melt the needle.

Many studies have addressed the nature of insight in scientific discovery. For example, Gestalt psychologists suggest that insight occurs when problem solvers see the original problem from a fresh perspective (Mayer, 1995). Other researchers have emphasized that the complexity of searching in a problem space has more to do with the structure of a problem space than the searcher (Perkins, 1995; Simon, 1981). In particular, Perkins distinguished two types of problem spaces. In a Homing Space, there are many clues and signposts such that navigating in such spaces is relatively easy. In contrast, a Klondike Space has very few such clues. The sparseness of clues is illustrated by Perkins (p. 498) in a widely known case of sudden insight — Charles Darwin's discovery of the principle of natural selection. According to Darwin's autobiography, in October 1838, he conceived the principle while he "happened to read for amusement 'Malthus on Population.' What is remarkable is that the next person arrived at the same principle 20 years later. What is even more remarkable is that the person, Alfred Russell Wallace, came up with the same idea while reading the same 1826 book by Malthus!

How could one increase the odds of stumbling on such clues? It becomes clear, from Sandstrom's notion of bibliographic microhabitats to Perkins' characterizations of Homing and Klondike spaces, that finding and recognizing clues is essential for both information foragers and problem solvers. Research in the data mining community on interestingness is particularly relevant (Hilderman & Hamilton, 2001; Liqiang & Howard, 2006). Interestingness is a quantitative measure of where a set of scientific ideas fit on the spectrum which ranges from the practice of normal science to that of paradigm-shifting ideas (Davis, 1971b). In this regard, interestingness lies between order and complete randomness, or chaos. We posit that three distinct ranges of scientific reports and ideas are those which are 1) either confirmatory or boring — there is nothing new for the scientific reader; the previously stated hypotheses have not been falsified yet, and are less and less likely to be so determined; 2) interesting, which may deny widely accepted assumptions, state new relationships between old ideas, propose new mechanisms, but do not require the reader to adopt wholly new ways of thinking; and 3) paradigm shifts and transformative discoveries. Interesting ideas are enlightening and surprising in a non-threatening way; in fact, a surprise is generally a pleasant one, in contrast to the experience of living through a shift of paradigm, especially when one's accepted paradigm is being replaced by a more successful one.

5.3.2 Literature-Based Discovery

Swanson and his colleagues pioneered a literature-based discovery approach to identify potentially valuable hypotheses (Swanson, 1986a, 1986b; Swanson, 1987; Swanson & Smalheiser, 1999). In essence, according to Swanson, the model of discovery from public knowledge is the A-B-C model, where the connections of A-B and B-C are known, but the connection of A-C is unknown. Thus A-C has the potential to become a candidate hypothesis for domain experts to evaluate. Using this template, a series of such candidate associations have been identified, including the connections between fish oil and Raynaud's syndrome (Swanson, 1986a), magnesium and migraine (Swanson, 1988), indomethacin and Alzheimer's disease (Smalheiser & Swanson, 1996).

Many researchers have subsequently adapted and refined Swanson's techniques. For example, Gordon and Lindsay conducted experiments with the MEDLINE medical literature database and extended the work of Swanson (Gordon & Lindsay, 1996; Lindsay & Gordon, 1999). They used lexical statistics to discover hidden connections in the medical literature. They argued that hidden connections are those that are unlikely to be found by examination of bibliographic citations or the use of standard indexing methods and yet establish a relationship between topics that might profitably be explored by scientific research. They also stressed that literature-based discovery cannot replace traditional empirical scientific research or even literature search, but rather supports them by providing the scientist with a means to organize more easily a potentially overwhelming amount of information.

Recently, Kostoff and his colleagues published a series of studies of literature-based discovery. These special studies presented a comprehensive approach for systematic acceleration of discovery and innovation, and demonstrated the generation of large amounts of potential discovery through five case studies describing the application of literature-based discovery to Raynaud's syndromes, cataracts, Parkinson's disease, multiple sclerosis, and water purification. He described the lessons learned from each application, and how the techniques can be improved further (Kostoff, 2008).

Where can we go from here? How often could a Nobel Prize award be characterized in terms of this A-B-C pattern of transitivity? Are there other patterns of scientific discoveries? If literature-based discovery is a computer-aided search in a problem space, what would it miss?

5.3.3 Spanning Diverse Perspectives

Effective strategies for making scientific discoveries have highlighted the ability to think creatively and look at a problem from a fresh perspective. Dunbar, for example, compared two different strategies of hypothesis generation using a Nobel Prize winning discovery as the test case (Dunbar, 1993). He found that it is a more effective discovery strategy to encourage researchers to consider novel alternative hypotheses. A 1992 special issue of *Theoretical Medicine* examined the mechanisms of scientific revolution and how the Nobel Prize committee selected scientific discoveries (Lindahal, 1992).

A longitudinal study of highly creative scientists in nano science and technology has found that it is not only the sheer quantity of publications that enables scientists to produce creative work but also their ability to effectively communicate with otherwise disconnected peers and to address a broader work spectrum (Heinze & Bauer, 2007). Why is it possible that communicating with otherwise disconnected scientists can lead to more creative work? What can one do specifically to come up with novel alternative hypotheses? How do we think outside the box?

There are many philosophical theories of scientific change. Philosophers of science (Laudan et al., 1986) argue that it would be useful to compare rival theories of scientific change against the history of science. Proponents suggest that conjectures of philosophical theories should be organized into theses so that one can compare these theories in terms of individual theses. Laudan et al. recommended rephrasing Lakatos' research programme, Laudan's research tradition, and Kuhn's paradigm in terms of a more generic notion of guiding assumptions. A superior theory of scientific change would be the one that has the most matches from the historical data. This idea was later criticized by (Radder, 1997), suggesting that it was far too ambitious.

Our needs here are different. Our goal is not to evaluate the value of individual philosophical theories of scientific change. Rather, what we need is an explanatory theory that can clarify the underlying mechanisms of specific scientific discoveries. In addition, we need a theory that can be instrumental for quantitative studies of scientific change.

Kuhn's paradigm-shift model of scientific revolutions (Kuhn, 1962, 1970) is probably the most widely known theory. It describes how science advances through a path of normal science, crisis, revolution, and new normal science.

A revolution involves a shift of world views from an old paradigm to a new paradigm. The paradigm-shift model has drawn criticisms. Critics argue that scientific change is often a lengthy process instead of a swift change as the paradigm-shift model suggests.

Cognitive scientists consider scientific discovery in common with everyday problem solving (Simon, Langley, & Bradshaw, 1981b). In (Klahr & Simon, 1999), four approaches to research on scientific discovery were identified; namely, historical accounts of scientific discoveries, psychological experiments with nonscientists working on tasks related to scientific discoveries, direct observation of ongoing scientific laboratories, and computational modeling of scientific discovery processes — by viewing them through the lens of the theory of human problem solving. The authors then considered these types of studies against a list of evaluative criteria, such as face validity, fine or coarse-grained, new phenomena, rigor and precision, social and motivational factors.

Many scholars have studied information and discovery pathways. Henry Small presented a series of examples from the history of science in which discoveries can be modeled as navigation between pairs of established experimental or theoretical findings (Small, 2000). One of his examples was from atomic physics in early twentieth century. There was no direct connection between experimental evidence on the spectrum for atomic hydrogen and evidence for hydrogen's nuclear structure until Niels Bohr's 1913 model for the hydrogen atom using a quantum hypothesis. Similarly, the Müller-and-Bednorz discovery of superconductivity was also seen as creating a path between the field of superconductivity and a class of compounds previously not thought to be promising candidates for superconductivity (Holton, Chang, & Jurkowitz, 1996; Small, 2000).

We notice a recurring theme in the diverse conceptualizations of scientific change. That is, profound scientific change tends to be connected to a broad range of brokerage mechanisms. Burt's structural holes are found not only in social networks but also in associative networks of intellectual, semantic, and other types of interrelationships. Because information flow around a structural hole is limited by the topological structure, those who are in the brokerage positions inherit advantages from their positions in such networks. Furthermore, structural holes in intellectual and cognitive networks appear to be a vital source of inspiration and creativity. Creative scientists draw inspirations from other disciplines. Research has found that great philosophers tend to be the ones who stayed in touch with competing schools of philosophy (Guiffre, 1999). Creative scientists are the ones who have the ability to communicate effectively with otherwise disconnected peers (Heinze & Bauer, 2007). Scientists make extra efforts to maintain contacts with scientists in different fields (Crane, 1972). Therefore, we have reached our central premise: bridging structural holes in a knowledge space is a valuable and viable mechanism for understanding and arriving at transformative scientific discoveries.

5.3.4 Bridging Intellectual Structural Holes

Now we will review some of the major conceptualizations of scientific change in light of the theory of structural holes (Burt, 1992, 2004, 2005). The theory of structural holes was originally developed in the context of social networks. The theory provides a meaningful and indeed enlightening framework for explaining why structural holes in intellectual networks such as co-citation networks may play an essential role in scientific discovery. Although this new conceptualization goes beyond the original scope of Burt's theory, we still refer them as structural holes for simplicity.

According to a sociological theory of scientific change (Fuchs, 1993), scientific discoveries are driven by two social factors, namely, peer competition and mutual dependence. Scientists seek novel discoveries to stay ahead in the invisible competition with their peers. As we have learned from the large body of relevant literature, inspirations often rise when different ways of thinking interact with one another. Structural holes in this sense span across patches of knowledge at different levels of self-organized structures, ranging from areas of research, fields of study, to disciplines.

From the information foraging perspective, establishing conceptual linkages between disparate patches of knowledge is a high-risk and high-return action. On the one hand, adapting a theory or a method from a 'foreign' discipline is likely to ensure its novelty in the 'home' discipline. It is more likely for us to think 'outside the box' in such situations. On the other hand, the fact that ideas and inspirations have obviously worked in another domain will drastically reduce the perceived risk that scientists may have to bear. This combination appears to give the maximum profitability associated with a structural hole.

From a philosophy of science's point of view, focusing on a structural hole also makes sense. In terms of Kuhn's paradigm-shift model, a competing paradigm is more likely to come from an unexpected place than right from the center of the currently predominant paradigm.

5.4 An Explanatory and Computational Theory of Discovery

A recurring theme across a wide variety of studies of scientific discovery, scientific change, creativity, and insight is that many creative ideas and profound discoveries can be traced to the work of a generic class of brokerage mechanisms. Brokerage mechanisms are not only found in social networks, such as networks of collaborators and coauthors, but also in more abstract conceptual networks of scientific knowledge, such as co-citation networks. For example, brokerage mechanisms have been seen to establish a previously unexpected linkage between structures of knowledge, connect two or more previously disparate fields of study, or recognize a meaningful analogy between distinct theories or hypotheses. Our new theory of scientific discovery is built on this recurring theme.

5.4.1 Basic Elements of the Theory

As the first step towards an explanatory and computational theory of scientific discovery, we concentrate on transformative and revolutionary discoveries. Transformative discoveries represent fundamental and revolutionary scientific changes. The growing interest in cyber-enabled discovery, e-science, and e-social science underlines the importance of advancing our understanding of how science works and identifying recurring mechanisms of creativity and discovery (Shneiderman, 2002, 2007). Supporting more transformative research is of critical importance in the fast-paced, science and technology-intensive world (NSF, 2007).

The fundamental premise of our theory is that a transformative discovery is made when a novel connection is established between two or more previously disparate units of scientific knowledge. Disparate units of scientific knowledge may include unconnected theories in different disciplines, isolated observations in the same field, or publications that have never been thought to be related. This conception is related to a number of approaches in the literature.

First, the brokerage-focused theory is inspired by the structural-hole theory of social networks (Burt, 2005). Furthermore, our theory adapts the brokerage mechanism and introduces it as a generic discovery mechanism for a wide variety of networks of scientific knowledge, such as citation networks, co-citation networks, networks of collaborating scientists, and other associative networks. The hypothesis that brokerage leads to greater collaborative creativity was tested in a study of collaborative inventors of utility patents (Fleming, Mingo, & Chen, 2007). Fleming et al. demonstrated that cohesive networks hamper creativity but aid in its transfer, particularly if the knowledge is complex and tacit. They tested more specific hypotheses such as a person is more likely to create new combinations if he or she brokers relations between otherwise disconnected collaborators. New combinations, as integrative work, are defined as a mechanism of creativity. In contrast, our theory focuses on transformative discoveries, which are conceptually more complex than new combinations of existing discoveries. For example, transformative discoveries often introduce new concepts and theories before integrative work becomes possible. The brokerage view also provides a simple explanation of why communicating with otherwise disconnected peer scientists is a distinct character of creative scientists (Heinze & Bauer, 2007).

Second, our theory is also related to literature-based discovery in that it shares the general goal of finding generative mechanisms of discovery. On the other hand, it differs from Swanson's famous A-B-C model. Instead of searching for a transitive closure of A→C, given A→B and B→C, we focus on the brokerage mechanism of discovery, which aims to establish an innovative connection between A and C. Another important difference is that we utilize structural properties of a network, whereas such properties are not used in Swanson's approach and its variations.

Third, our theory is related to network evolution models in complex network analysis. Preferential attachment models, for example, characterize the growth of a network in a process that popular nodes will become even more popular as new nodes and links are added to the network (Albert & Barabási, 2002; Barabási et al., 2002). The popularity of a node can be broadly defined by an attribute function of node, such as prestige, age, or by other ranking mechanisms. Such processes often result in scale-free networks, which are characterized by power law distributions of node degrees. While earlier preferential attachment models assume that each new coming node is fully aware of the prestigious status of every existing node, more recent studies have relaxed the assumption to ranking functions defined on a subset of the existing nodes instead (Fortunato, Flammini, & Menczer, 2006). In contrast, the brokerage mechanism in our theory provides a growth mechanism by building connections across structural holes between two or more thematic networks. A brokerage-driven growth is distinct from growths that can be modeled by preferential attachment.

Fourth, our theory extends earlier efforts for predicting Nobel Prize winners based on citation ranking (Garfield, 1992). Thomson Reuters' Citation Laureates[1] are also in this category. Our approach is distinct in several important ways. Although using citation ranking alone has the advantage of simplicity, we take multiple factors such as structural holes and the rate of citation growth into account in order to better accommodate the complexity. In addition, we are concerned with the possibly delayed identification due to the time taken for the citation profile of a scholarly publication to become prominent enough to be noticed. We expect that using structural properties in the theory can resolve the issue to some extent.

Fifth, our theory provides an explanatory mechanism for the diffusion process associated with a transformative discovery. Once a brokerage connection is established between previously disparate areas, it would facilitate the information flow between these areas. In other words, we expect that the newly discovered connection will accelerate the diffusion process. Interestingly, the expected effect on diffusion can be explained in terms of the information foraging theory (Pirolli, 2007). According to the information foraging theory, searchers need to evaluate multiple patches of information. They need to make decisions on which patch they should focus and how long they should spend on a patch before they move on. Their decisions are essentially determined by the perceived profitability of each move. The higher the perceived

[1] http://scientific.thomsonreuters.com/nobel/nominees/

profitability, the more likely they will decide to go ahead and take the action. The newly discovered connection will increase the perceived profitability because the discovery not only reduces the risk, but also provides concrete and positive examples of success. Therefore, we could conjecture that the increased perceived profitability will be translated into bursts of observed frequencies such as citation and co-citation counts.

Finally, the theory is related to but distinct from the notion of co-citation pathways through science (Small, 2000). The creation of co-citation pathways aims to traverse scientific literature through a chain of highly co-cited pairs of papers. Small found a co-citation pathway of 331 highly cited documents starting from economics and ending in astrophysics (Small, 1999). He observed that each successive document in this path embodies an information transition towards the destination topic and, in most cases, such transitions are surprisingly smooth. In contrast, the focus of our theory is on novel connections that bridge previously disparate fields. Although in theory such connections may appear as part of a co-citation pathway, it seems to be more likely that brokerage connections would either deviate from pathways of highly co-cited documents or not be selected altogether because of a high co-citation threshold. Nevertheless, more investigations are needed to clarify the relationships in detail.

5.4.2 Structural and Temporal Properties

Now we will focus on two specific properties of scientific discoveries that can be derived from our theory, namely a structural property of a discovery in a network of scientific papers measured by the betweenness centrality (Freeman, 1977) and a temporal property measured by burstness of citations (Kleinberg, 2002).

Our theory states that a transformative discovery is made when a bridging connection is established between two or more previously disconnected patches of knowledge. If we represent knowledge in the form of networks, such bridging connections would be links between two or more disconnected networks or components of a network. Such connections in networks can be computationally identified using the betweenness centrality. In fact, one can even compute the would-be centrality of a node if it were to have some of the non-existent connections. The betweenness centrality of a particular node or link measures the importance of the node or link in connecting any two nodes in the network. A node or link that is essential for linking many pairs of nodes will have a high betweenness centrality. Therefore, a paper with a high betweenness centrality is potentially a transformative discovery. In addition, it would be possible to use this metric to identify potential future discoveries by calculating the would-be betweenness centrality of a hypothetical connection between two disparate areas of existing knowledge networks.

It is possible to devise computer simulation algorithms to identify a short list of such candidates of discovery to be made.

Several relevant concepts have been derived from betweenness centrality metrics. For example, in CiteSpace (Chen, 2004; Chen, 2006), pivotal points in co-citation networks are identified based on their betweenness centrality. These are the points that are cited with different co-citation clusters. We have mentioned earlier that co-citation clusters correspond to thematic structures. Therefore, points connecting different thematic structures are candidates of intellectual turning points.

In a journal co-citation network, high betweenness centrality is an indicator of interdisciplinary journals (Leydesdorff, 2007). Taken together, it suggests that the betweenness centrality indicator can be used at various scales of granularity to indicate and predict transformative changes. Furthermore, betweenness centrality is found to correlate with long-term citations predicted into the future (Shibata, Kajikawa, & Matsushima, 2007). This finding would be consistent with our conceptualization of scientific discovery in that scientists will pay constant attention to structural holes for future discoveries.

The emphasis on betweenness centrality differentiates our theory from other approaches to network evolution models, especially preferential attachment models. Instead of adding one link at a time to the most prominent node in a given network, our theory says that a scientific discovery needs to form a path spanning over an intellectual structural hole. As a result, the newly added scientific discovery would have a high betweenness centrality. Our theory also implies that a node with high betweenness centrality would be more valuable to a foraging scientist than a node with a higher citation count but lower betweenness centrality. While the latter may bring nothing new to a scientist who is well aware of the highly cited work, the former may lead to new insights that a scientist may actually act on. Thus, betweenness centrality can be translated into interestingness, which can be in turn translated into actionability. We have indeed observed in our previous work that the most cited references are not necessarily the most revolutionary ones (Chen, 2004; Chen & Kuljis, 2003).

Betweenness centrality is a structural property of a network. Our theory also leads to temporal properties of an evolving network, for example, the burstness of citation of a reference over time. Burst detection is a class of algorithms to identify changes of a variable over a period of time with reference to others in the same population (Kleinberg, 2002). Our theory suggests that a burst of citation could be a good indicator of a transformative discovery, especially from a profitability-guided foraging point of view, when it is observed with a structural property such as the betweenness centrality metric. As we have analyzed earlier, a brokerage discovery would increase the perceived profitability for moving from one patch of knowledge to another. As a result, the increased profitability and reduced risks should boost the adaptation and diffusion of the new discovery.

Would the absence of such structural and temporal properties rule out the possibility of a transformative discovery? This issue is concerned with the scope of the theory. Further investigations are needed. In the following section, we present some examples to further clarify the major properties derived from the theory.

5.4.3 Integration

We focus on cases in which both structural and temporal properties are observed and evaluate the role of brokerage mechanisms in such cases. In addition to study individual properties such as the betweenness centrality, the burst rate, or citation counts, we introduce a group of generic metrics $\sigma_n(v, G, T, \rho_1, \rho_2, \ldots, \rho_n)$ as indicators of the potential transformative strength of a node v in a given network G over a time interval T with respect to n properties. Each ρ_i is a function $\rho_i(v, G, T)$ in the range of [0, 1]. These metrics can be generically defined as the geometric mean of multiple normalized properties $\rho_1, \rho_2, \ldots, \rho_n$ in the range of [0, 1]. The maximum possible value of σ is 1 when all the individual properties have the maximum value of 1. The minimum possible value of σ is 0 when any of the individual properties is 0.

$$\sigma_n(v, G, T, \rho_1, \ldots, \rho_n) = \left(\prod_{i=1}^{n} \rho_i\right)^{\frac{1}{n}} \tag{1}$$

In particular, in the following case studies, the metric σ is defined based on ρ_{citation}, $\rho_{\text{centrality}}$, ρ_{burst} as follows. The definitions of $\rho_{\text{centrality}}$ and ρ_{burst} can be found in (Brandes, 2001; Freeman, 1977; Kleinberg, 2002).

$$\sigma_1(v, G, T, \rho_{\text{burst}}) = \left(\prod_{i=\text{burst}} \rho_i\right)^{\frac{1}{1}} = \rho_{\text{burst}} \tag{2}$$

$$\sigma_1(v, G, T, \rho_{\text{centrality}}) = \left(\prod_{i=\text{centrality}} \rho_i\right)^{\frac{1}{1}} = \rho_{\text{centrality}} \tag{3}$$

$$\sigma_1(v, G, T, \rho_{\text{citation}}) = \left(\prod_{i=\text{citation}} \rho_i\right)^{\frac{1}{1}} = \rho_{\text{citation}} \tag{4}$$

$$\begin{aligned}\sigma_2(v, G, T, \rho_{\text{burst}}, \rho_{\text{centrality}}) &= \left(\prod_{i=\text{burst},\text{centrality}} \rho_i\right)^{\frac{1}{2}} \\ &= \sqrt{\rho_{\text{burst}} \cdot \rho_{\text{centrality}}}\end{aligned} \tag{5}$$

$$\sigma_3(v, G, T, \rho_{\text{burst}}, \rho_{\text{centrality}}, \rho_{\text{citation}}) = \left(\prod_{i=\text{burst,centrality,citation}} \rho_i \right)^{\frac{1}{3}}$$

$$= \sqrt[3]{\rho_{\text{burst}} \cdot \rho_{\text{centrality}} \cdot \rho_{\text{citation}}} \qquad (6)$$

Note that $\sigma_1(\rho_{\text{citation}})$, a special case of the generic definition, ranks the significance of a reference based on its citations as seen in earlier efforts for predicting Nobel Prize winners based on citation counts (Garfield, 1992). We will also compare pair-wise Pearson correlation coefficients between σ_1, σ_2 and σ_3 indices of centrality, burst, and citation frequency in order to identify the simplest and effective metrics among them.

In summary, our theory suggests that σ indices would be a good indicator of potential transformative discoveries. Furthermore, once a reference is identified with a high σ index, the theory provides an explanatory framework such that we can focus on the precise brokerage connections at work. The theory also suggests alternative ways to model the evolution of a network by taking brokerage connections into account. According to our theory, a subset of Nobel Prize discoveries will be transformative discoveries. More transformative discoveries would be expected from the recipients of a variety of other awards in science. In addition, we expect that transformative discoveries can be identified by these σ metrics at an earlier stage than by single-dimensional ranking systems. In terms of diffusion, we expect that transformative discoveries in general will lead to a more rapid and sustained diffusion process. If we see the diffusion process as an information foraging process by the scientific community as a whole, transformative discoveries, i.e., brokerage connections across structural holes, would have a higher perceived profitability, which would motivate and stimulate the diffusion process. It also follows that the domain-wide foraging process will spend more time with transformative discoveries than other patches of scientific knowledge.

5.4.4 Case Studies

In each case study, CiteSpace (Chen, 2006) was used to construct a co-citation network of the references relevant to the chosen topic. We followed the general procedure described in (Chen, 2004; Chen, 2006). Bibliographic records were retrieved from the Web of Science with a topical search for articles only. Reviews, editorials, and other document types were excluded from the analysis.

CiteSpace uses a time-slicing mechanism to generate a synthesized panoramic network visualization based on a series of snapshots of the evolving network across consecutive time slices[2]. Each node in the network represents a reference cited by records in the retrieved dataset. A line connecting two

[2] http://cluster.cis.drexel.edu/∼cchen/citespace/

nodes represents one or more co-citation instances involving the two references. Colors of co-citation links correspond to the earliest year in which co-citation associations were first made. Each node is shown with a tree-ring of citation history in the same color scheme, representing the history of citations received by the underlying reference.

Structural-hole and burst properties are depicted in two distinct colors — purple and red — in visualizations. If a node is rendered with a purple ring, it means it has a strong betweenness centrality. The purple color can only appear as the color of the outermost rim of a node. The thickness of the purple ring is proportional to the degree of the centrality: the thicker, the stronger the betweenness centrality. In contrast, if a node has red rings, these red rings represent the presence and strength of its burst property. It can appear as the color of any inner rings of the tree ring of a node. The presence of one or more red rings on a node indicates a significant citation burst was detected. In other words, there was a period of time in which citations to the reference increased sharply with respect to other references in the pool, hence the name CiteSpace.

5.4.4.1 Peptic Ulcer

The Nobel Prize in Physiology or Medicine for 2005 was awarded jointly to Barry J. Marshall and J. Robin Warren for their discovery of "the bacterium Helicobacter pylori and its role in gastritis and peptic ulcer disease." We choose *peptic ulcer* as the topic area.

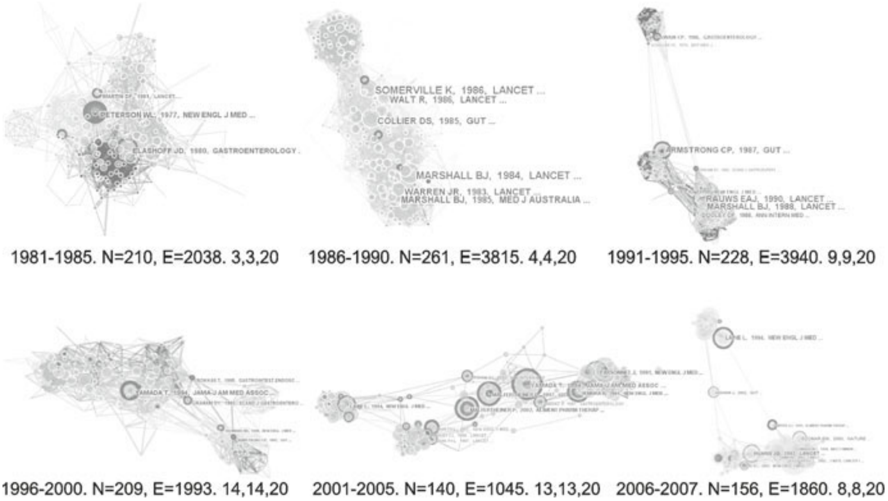
According to Marshall's Nobel Prize lecture (Marshall, 2005), Marshall and Warren conducted a study in the 1980s and found 100% of 13 patients with duodenal ulcer were infected by Helicobacter pylori. They discovered that peptic ulcer was caused by a bacterial infection, unlike the then predominant understanding that ulcers were caused by other reasons such as stress and acid in the stomach. The discovery established that very young children acquired the Helicobacter organism, a chronic infection which caused a lifelong susceptibility to peptic ulcers. Helicobacter was generally accepted after 1994 as the cause of most gastroduodenal diseases including peptic ulcer and gastric cancer.

We analyzed a co-citation network of peptic ulcer research to identify structural and temporal properties associated with the Helicobacter pylori discovery. Bibliographic records on peptic ulcer between 1980 and 2007 were retrieved from the Web of Science with a topic search for 'peptic ulcer'. CiteSpace was used to construct a co-citation network of peptic ulcer research between 1980 and 2007.

Fig. 5.6 shows a series of 5-year snapshots of the co-citation network as it evolved over time. In each diagram, five colors match to the five years in the order of blue, cyan, green, yellow, and orange. Thus, an orange cluster would be formed in the 5th year of a given 5-year interval. For example, a node with essentially a green tree-ring means the reference was mostly cited in the 3rd year of the time interval.

The captions below network snapshots record the time interval, the number of nodes, the number of co-citation links, and three thresholds. For example, the caption "1981 – 1985. N=210, E=2038. 3,3,20" under the first snapshot of the network means that the network was formed between 1981 and 1985, consisting of 210 references and 2,038 co-citation pairs. Each reference has received at least 3 citations in one of the 5 years during this period.

According to independent sources (Pincock, 2005), the first major publication of the Helicobacter pylori discovery was (Marshall & Warren, 1984). Marshall-1984 appeared in the 1986 – 1990 network with essentially cyan and green citation rings, which means it received its citations mostly in 1987 and 1988. It is quite possible that Marshall-1984 was cited as soon as it was published in the 1981 – 1985 time interval, but it did not reach the top of the most cited list until the 1986 – 1990 network. The six snapshots also demonstrate that peptic ulcer research has evolved constantly with new references reaching the top cited levels.


MARTIN CP, 1981, LANCET ...
PETERSON WL, 1977, NEW ENGL J MED ...
ELASHOFF JD, 1980, GASTROENTEROLOGY ...
SOMERVILLE K, 1986, LANCET ...
WALT R, 1986, LANCET ...
COLLIER DS, 1985, GUT ...
MARSHALL BJ, 1984, LANCET ...
WARREN JR, 1983, LANCET ...
MARSHALL BJ, 1985, MED J AUSTRALIA ...
ARMSTRONG CP, 1987, GUT ...
RAUWS EAJ, 1990, LANCET ...
MARSHALL BJ, 1988, LANCET ...
YAMADA T, 1994, JAMA-J AM MED ASSOC ...
YAMADA T, 1994, JAMA-J AM MED ASSOC ...
LAINE L, 1994, NEW ENGL J MED ...

1981-1985. N=210, E=2038. 3,3,20

1986-1990. N=261, E=3815. 4,4,20

1991-1995. N=228, E=3940. 9,9,20

1996-2000. N=209, E=1993. 14,14,20

2001-2005. N=140, E=1045. 13,13,20

2006-2007. N=156, E=1860. 8,8,20

Fig. 5.6 A co-citation network of references on peptic ulcer research (1980 – 1990). Source: (Chen, Chen, Horowitz, Hou, Liu, & Pellegrino, 2009). (see color figure at the end of this book)

Fig. 5.7 shows a panorama view of the entire time interval of the dataset (1980 – 2007). Marshall-1984 has a prominent structural property — a high betweenness centrality (a large purple ring). Although it does demonstrate a temporal property of burstness, its burst rate is detectable but not as strong as some of its neighbors. The burst period was between 1986 and 1988, which is consistent with our observations in the earlier 5-year snapshot series. The overview network shows that Marshall-1984 is in a dense cluster with numerous references with citation bursts, suggesting other high-impact references were present in the landscape of peptic ulcer research.

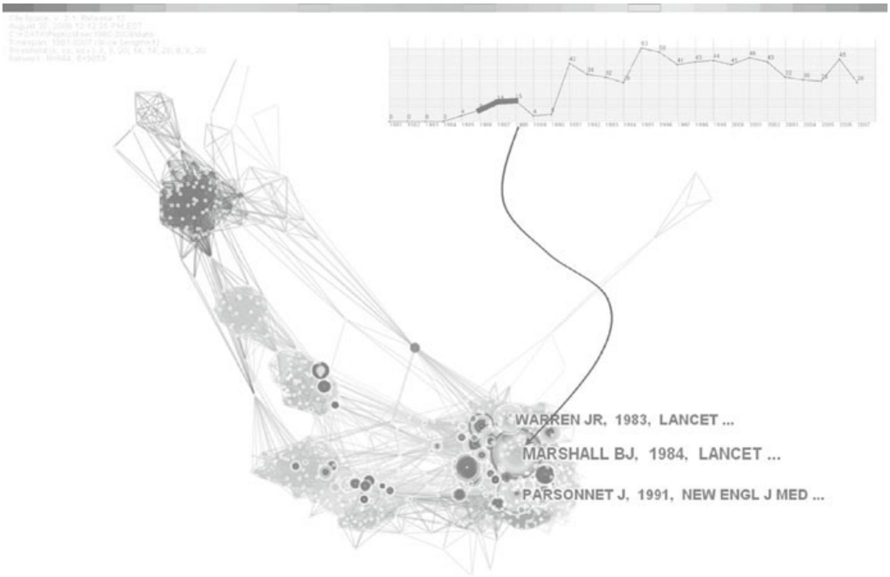

WARREN JR, 1983, LANCET ...
MARSHALL BJ, 1984, LANCET ...
PARSONNET J, 1991, NEW ENGL J MED ...

Fig. 5.7 A co-citation network of references cited between 1981 and 2007 in peptic ulcer research. Source: (Chen et al., 2009). (see color figure at the end of this book)

As shown in Table 5.3, Marshall-1984 was the most cited reference (711 citations) and the highest betweenness centrality ($\rho_{\text{centrality}}$ of 0.393). On the other hand, its burst rate ranked the 372nd. Marshall and Warren encountered resistances in getting their discovery accepted by the peptic ulcer research community. The slow acceptance was documented (Pincock, 2005), which may in part explain its relatively low burst rate. In contrast, Marshall-1988 has the highest σ_2 of 0.416. It was entitled *Prospective double-blind trial of duodenal ulcer relapse after eradication of Campylobacter pylori.* In his Nobel Prize lecture, Marshall dated the acceptance of his work as the 1994 NIN consensus conference in Washington DC.

The last column in Table 5.3 contains the σ_2 index, i.e., the geometric mean of the burst and centrality metrics. According to our theory, a transformative discovery is a brokerage between previously disconnected areas of scientific knowledge. The σ_2 index takes into account both structural and temporal properties that a discovery over a structural hole would demonstrate. In this case, Marshall-1988 was the highest ranking candidate according to the σ_2 index, despite its citation count of 421 was much less than Marshall-1984. Validating the true value of Marshall-1988 is beyond our own expertise and beyond the scope of the analysis. Properly validating the value of references with such strong combinations of structural and temporal properties will be an important issue to be addressed in the future work of our construction of the theory. It is also related to the potential power of predicting high-impact discoveries even before it reaches its citation peaks or while they are overshadowed by other highly cited references.

Table 5.3 Top 5 most cited references in peptic ulcer research (1980 – 2007).

Citation	Author	Year	Source	Vol.	Page	ρ_{burst}	$\rho_{\mathrm{centrality}}$	σ_2
711	MARSHALL BJ	1984	LANCET	1	1311	0.138	0.393	0.232
581	PARSONNET J	1991	NEW ENGL J MED	325	1127	0.208	0.143	0.172
579	WARREN JR	1983	LANCET	1	1273	0.165	0.250	0.203
466	YAMADA T	1994	JAMA	272	65	0.635	0.071	0.213
421	MARSHALL BJ	1988	LANCET	2	1437	0.607	0.286	0.416

5.4.4.2 Gene Targeting and the Sticky Effect

The Nobel Prize in Physiology or Medicine for 2007 was awarded jointly to Mario R. Capecchi, Martin J. Evans and Oliver Smithies for their discoveries of "principles for introducing specific gene modifications in mice by the use of embryonic stem cells." This field of study is often known as gene targeting. We applied the same procedure described earlier for gene targeting. We used topic searches in the Web of Science for 'gene target*', 'genetic* target*', and 'gene* knock*' for genetic knock-out, another term used to describe the techniques in general. A total of 8,160 bibliographic records were retrieved between 1985 and 2007.

Fig. 5.8 shows an overview of a co-citation network of gene targeting references cited between 1985 and 2007. Notably, the three nodes with the highest betweenness centrality scores are all connected to the 2007 Nobel Prize awards: Capecchi-1989, Mansour-1988, and Thomas-1987. Here only the first author of each paper was recorded in the Web of Science cited reference field. The three papers represent a series of innovations of fundamental techniques for gene targeting. Unlike the case with Marshall-1984, all three groundbreaking gene targeting papers have strong citation bursts, shown in Fig. 5.7 as the thickened rising curves. It also becomes clear that these curves have subsequently peaked and steadily declined, which means they are getting fewer and fewer citations. The visualization confirms this pattern. The network shows that the most recent active areas are located in the lower left quadrant of the visualization.

The 2007 Nobel Prize awards mentioned the use of embryonic stem cells. Techniques developed in embryonic stem cell research turned out to be critical to the gene targeting techniques. Martin J. Evans, who shared the 2007 Nobel Prize, is known as the architect of embryonic stem cells. The pioneering discovery made by Evans in 1981 (Evans & Kaufman, 1981) was in fact cited by the Thomas-1987 gene targeting paper. Evans-1981 was cited 1,681 times in the Web of Science, although it was not highly cited within the gene targeting dataset we analyzed. Techniques developed by Evans were among the many building blocks that were necessary for the ultimate gene

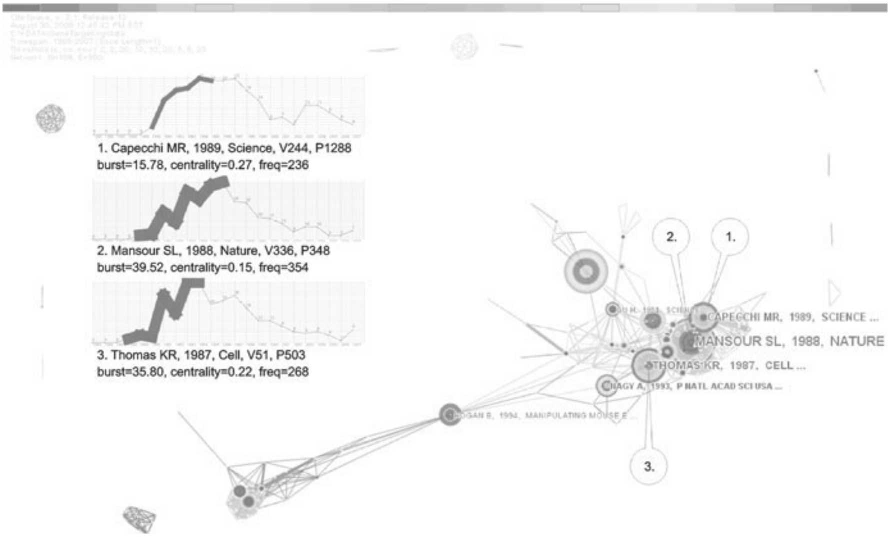


Fig. 5.8 A co-citation network of references cited between 1985 and 2007 in gene targeting research. References with the strongest betweenness centrality scores are labeled. The burst periods of their citations are shown as the thickened curves in the three diagrams to the left. Source: (Chen et al., 2009). (see color figure at the end of this book)

targeting techniques. A number of questions can be addressed from our theory of discovery. For example, how easy or how hard was it to discover Evans-1981 for the needs of gene targeting? Who were the first citers of Evans-1981. What was Evans' own research field and how was it related to gene targeting? What are the other building blocks used by these Nobel laureates in their discoveries? Were their discoveries taking place over an intellectual structural hole? How did their discovery change the association between existing intellectual structures?

Table 5.4 lists the top 5 references by σ_2 — the geometric mean of $\rho_{\text{centrality}}$

Table 5.4 Top 5 references by σ_2 — the geometric mean of centrality and burstness.

Author	Year	Source	Vol.	Page	Citations	ρ_{burst}	$\rho_{\text{centrality}}$	σ_2
THOMAS KR	1987	CELL	51	503	268	0.851	0.537	0.676
HOGAN B	1994	MANIPULATING MOUSE E	BOOK		136	0.409	1.000	0.639
MANSOUR SL	1988	NATURE	336	348	354	0.940	0.366	0.586
CAPECCHI MR	1989	SCIENCE	244	1288	236	0.375	0.659	0.497
NAGY A	1993	P NATL ACAD SCI USA	90	8424	182	0.346	0.463	0.400

and ρ_{burst}. The 1st, 3rd, and 4th references are connected to the Nobel Prize winning discoveries. Note that the first discovery paper Thomas-1987 has the highest ranking although its citation count of 268 is not the highest. The 2nd reference is a book. If we consider journal articles only, the first three references would be all related to the Nobel discoveries (see Fig. 5.9).

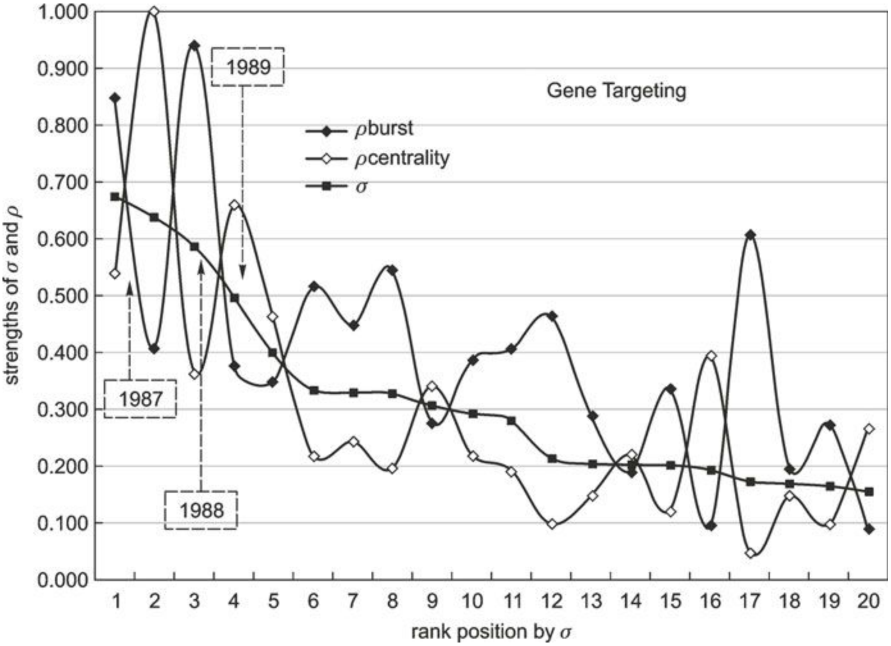


Fig. 5.9 Nobel Prize winning papers are ranked among the highest by the σ_2 index. Source: (Chen et al., 2009).

Fig. 5.10 is a visualization of the areas associated with the Nobel Prize winning discoveries in gene targeting research. The visualization was generated based on citing articles with 15 or more citations in the Web of Science. In other words, these citing articles themselves have made impacts on the field in their own right. Co-cited references are aggregated into clusters. The diffusion of knowledge is tracked by showing how co-citation footprints move from one cluster to another over time and how long they stay in particular clusters. The history of the evolution can be seen as an information foraging process participated in by all the scientists in the field. For example, the *embryo-derived stem cell* (cluster #11) attracted a lot of citations in 1987 (shown as a high density co-citation cluster in red). In 1988, the foraging process moved to *DNA delivery method* (cluster #19) above cluster #11. All three papers associated with the 2007 Nobel Prize are concentrated in cluster #12 — *gene correction*. During 1989 and 1990, much of the foraging process was inside cluster #12. We also studied the diffusion process over a longer period of time and the foraging process appeared to spend much longer time with cluster #12 than any other clusters. Our general hypothesis is that transformative discoveries tend to retain the foraging process longer than other patches of knowledge. Further investigations are needed. The connection between structural-hole theory and information foraging theory is an important research direction for further investigation.

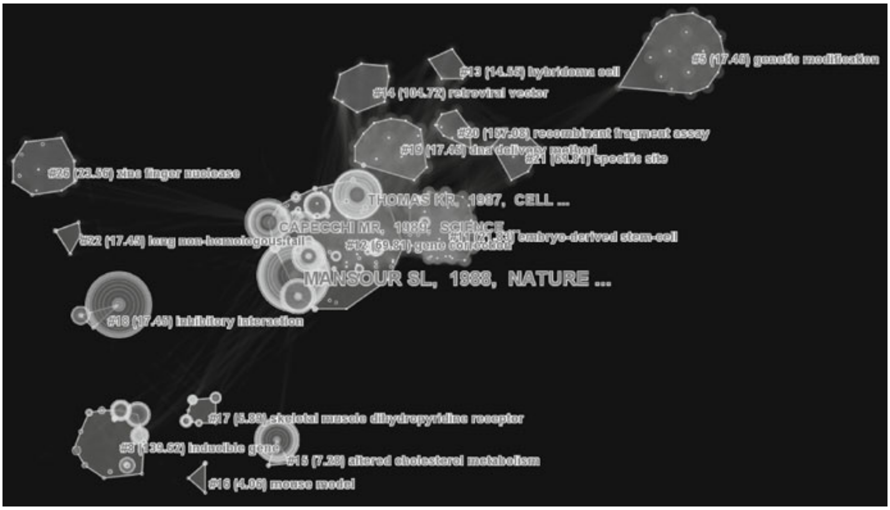
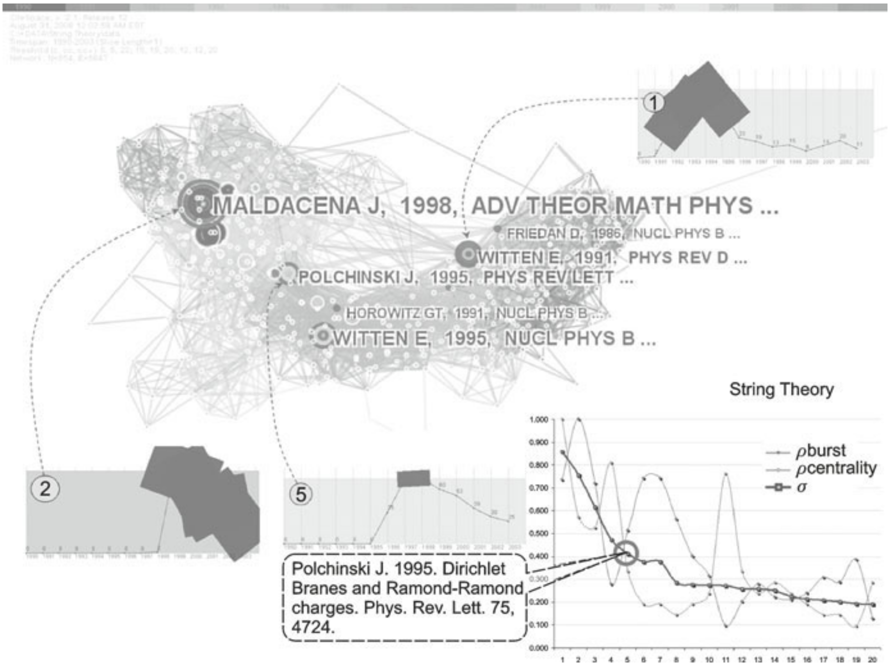

#5 (17.45) genetic modification
#13 (14.55) hybridoma cell
#14 (104.72) retroviral vector
#20 (157.08) recombinant fragment assay
#19 (17.45) dna delivery method
#21 (69.81) specific site
#26 (23.56) zinc finger nuclease
THOMAS KR, 1987, CELL ...
CAPECCHI MR, 1989, SCIENCE
#11 (21.83) embryo-derived stem-cell
#22 (17.45) long non-homologous tail
#12 (69.81) gene correction
MANSOUR SL, 1988, NATURE ...
#18 (17.45) inhibitory interaction
#17 (5.89) skeletal muscle dihydropyridine receptor
#8 (139.62) inducible gene
#15 (7.28) altered cholesterol metabolism
#16 (4.06) mouse model

Fig. 5.10 A diffusion map of gene targeting research between 1985 and 2007. Selection criteria are at least 15 citations for citing articles and top 30 cited articles per time slice. Polygons represent clusters of co-cited papers. Each cluster is labeled by title phrases selected from papers citing the cluster. Red lines depict co-citations made in the current year. The concentrations of red lines track the context in which co-citation clusters are referenced. Source: (Chen et al., 2009). (see color figure at the end of this book)

5.4.4.3 String Theory

The third illustrative example is string theory in physics (Schwarz, 1982). We have studied this topic as an example of Kuhn's scientific revolutions (Chen, 2004; Chen & Kuljis, 2003). According to (Schwarz, 1982), two conceptual revolutions occurred in string theory: one was in 1980s and the other in 1990s. Using relevant citation records between 1990 and 2003, we conducted a similar study of string theory and focused on the two properties of the revolutionary papers for the second string theory revolution.

Fig. 5.11 shows an overview of a visualized co-citation network of references in the period of 1990 – 2003. According to Schwarz (1982), Polchinski-1995 marked the second string theory revolution. Polchinkski-1995 is ranked the 5th by the geometric mean index. The visualization shows it has a relatively strong betweenness centrality and its burst rate is not as prominent as a few others in the field. Witten-1991 has the highest geometric mean index ranking, followed by Maldacena-1998; both have shown strong betweenness centrality and burstness.



Polchinski J. 1995. Dirichlet Branes and Ramond-Ramond charges. Phys. Rev. Lett. 75, 4724.

Fig. 5.11 A co-citation network of references cited between 1990 and 2003 in string theory. Polchinski-1995 marked the beginning of the second string theory revolution. Maldacena-1998 is highly transformative and brokerage link between string theory and particle theories. The three embedded plots show the burst periods of citations of Witten-1991, Maldacena-1998, and Polchinski-1995. Source: (Chen et al., 2009). (see color figure at the end of this book)

Maldacena-1998 is not only strong in both centrality and burstness, it is also the most cited reference in this dataset. We contacted Juan Maldacena directly and asked him to identify the nature of his major contributions in this article to String Theory. The transformative nature is evident in his reply: "It connected two different kinds of theories: 1) particle theories or gauge theories and 2) string theory. Many of the papers on string dualities (and this is one of them) connect different theories. This one connects string theory to more conventional particle theories." Maldacena's contribution is highlighted on the TIME 100 Innovator website[3] as "he forged a connection between the esoteric formulas of string theory and the rest of mainstream physics." Even more intriguingly from the perspective of our brokerage theory, he "has been able to suggest a way to knit together two theories previously thought to be incompatible: quantum mechanics, which deals with the universe at its

[3] http://www.time.com/time/innovators/science/profile_maldacena.html

smallest scales; and Einstein's general theory of relativity, which deals with the very largest." In addition, our search on the web reveals that he is the recipient of the 2007 Dannie Heineman Prize for Mathematical Physics[4] "for profound developments in Mathematical Physics that have illuminated interconnections and launched major research areas in Quantum Field Theory, String Theory, and Gravity."

Table 5.5 shows pair-wise Pearson correlation coefficients between normalized burst and centrality scores, the σ_2 index of burst and centrality, and the σ_3 index of burst, centrality, and citation frequency. The σ_2 and σ_3 indices are strongly correlated ($r = 0.9780$), suggesting that, at least in this case, the σ_3 index is redundant and we can simply focus on σ_2. The correlation coefficients also show that burstness and centrality are almost independent measures, although they both have some connections to citation counts. This is a simple justification of our choice to use both burstness and centrality to construct σ_2 as an index of high-impact discoveries. More comprehensive validations may consider other measures such as the h-index and its numerous variations, e.g. (Antonakis & Lalive, 2008; Hirsch, 2005b).

Table 5.5 Pearson correlation coefficients between individual and synthetic indices.

	ρ_{burst}	$\rho_{\text{centrality}}$	$\sigma_2(\rho_{\text{burst}}, \rho_{\text{centrality}})$
ρ_{citation}	0.8026	0.3618	
ρ_{burst}		0.0409	
$\sigma_3(\rho_{\text{burst}}, \rho_{\text{centrality}}, \rho_{\text{citation}})$			0.9780

5.5 Summary

In this chapter we have discussed an information-theoretic view of visual analytics as a general framework for sensemaking and analytic reasoning. The main thesis is that we need to maintain our situational awareness in light of new evidence. We have also introduced the notion of turning points as new information that can potentially change our mental models. A key conjecture we make based on existing studies of scientific discoveries is that there are generic mechanisms of discovery. We have discussed a few examples of generic mechanisms such as scientific discovery as a special case of problem solving, literature-based discovery that seeks new hypotheses based on missing links between disparate bodies of knowledge, and boundary spanning mechanisms such as the one derived from the structural hole theory.

Taken these together, we have introduced an explanatory and computational theory of scientific discovery. The theory provides an extensible framework, which currently consists of structural and temporal properties

[4] http://www.aps.org/programs/honors/prizes/prizerecipient.cfm?name=Juan%20 Maldacena&year=2007

as the necessary conditions for potentially significant discoveries. We have illustrated the potential of the theory through three case studies. The theory will be further developed in the next few chapters and used to derive metrics for identifying the potential of transformative research.

References

Adar, E., Zhang, L., Adamic, L.A., & Lukose, R.M. (2004). Implicit structure and the dynamics of blogspace. In Proceedings of the workshop on the weblogging ecosystem at 13th international world wide web conference.

Albert, R., & Barabasi, A. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics, 74(1), 47-97.

Anderson, T., Schum, D., & Twining, W. (2005). Analysis of evidence. (2nd ed.). Cambridge, England: Cambridge University Press.

Antonakis, J., & Lalive, R. (2008). Quantifying scholarly impact: IQp versus the Hirsch h Journal of the American Society for Information Science and Technology, 59(6), 956-969.

Barabási, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. Physica A, 311, 590-614.

Bartels, L. (1988). Issue voting under uncertainty: An empirical test. American Journal of Political Science, 30, 709-728.

Bederson, B.B., & Shneiderman, B. (2003). Theories for understanding information visualization. In the craft of information visualization: Readings and reflections (pp. 349-351): Morgan Kaufmann.

Bettencourt, L.M.A., Castillo-Chavez, C., Kaiser, D., & Wojick, D.E. (2006). Report for the office of scientific and technical information: Population modeling of the emergence and development of scientific fields.

Bettencourt, L.M.A., Kaiser, D.I., Kaur, J., Castillo-Chavez, C., & Wojick, D.E. (2008). Population modeling of the emergence and development of scientific fields. Scientometrics, 75(3), 495-518.

Bradshaw, G.F., Langley, P.W., & Simon, H.A. (1983). Studying scientific discovery by computer simulation. Science, 222(4627), 971-975.

Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 25(2), 163-177.

Brannigan, A., & Wanner, R.A. (1983). Historical distributions of multiple discoveries and theories of scientific change. Social Studies of Science, 13, 417-435.

Brody, T., Harnad, S., & Carr, L. (2006). Earlier web usage statistics as predictors of later citation impact. Journal of the American Association for Information Science and Technology, 57(8), 1060-1072.

Brush, S.G. (1994). Dynamics of theory change: The role of predictions. In Proceedings of the 1994 Biennial Meeting of the Philosophy of Science Association(pp. 133-145). East Lansing, MI.

Brush, S.G. (1995). Prediction and theory evaluation in physics and astronomy. In A.J. Kox & D.M. Siegel (Eds.), No Truth Except in the Details (pp. 299-318). Dordrecht: Kluwer Academic Publishers.

Burt, R.S. (1992). Structural holes: The social structure of competition. Cambridge, Massachusetts: Harvard University Press.

Burt, R.S. (2001). The social capital of structural holes. In N.F. Guillen, R. Collins, P. England & M. Meyer (Eds.), New directions in economic sociology. New York: Russell Sage Foundation.

Burt, R.S. (2004). Structural holes and good ideas. American Journal of Sociology, 110(2), 349-399.

Burt, R.S. (2005). Brokerage and closure: An introduction to social capital. Oxford, UK: Oxford University Press.

Cahn, R.W. (1970). Case histories of innovations. Nature, 225, 693-695.

Chen, C. (2003). Mapping scientific frontiers: The quest for knowledge visualization. London: Springer.

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. Proc. Natl. Acad. Sci. USA, 101(suppl), 5303-5310.

Chen, C. (2006). CiteSpace Ⅱ: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57(3), 359-377.

Chen, C. (2008). An information-theoretic view of visual analytics. IEEE Computer Graphics & Applications, 28(1), 18-23.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics, 3(3), 191-209.

Chen, C., & Kuljis, J. (2003). The rising landscape: A visual exploration of superstring revolutions in physics. Journal of the American Society for Information Science and Technology, 54(5), 435-446.

Chubin, D.E. (1976). The conceptualization of scientific specialties. The Sociological Quarterly, 17(4), 448-476.

Collins, R. (1998). The sociology of philosophies: A global theory of intellectual change. Cambridge, MA: Harvard University Press.

Crane, D. (1972). Invisible colleges: diffusion of knowledge in scientific communities. Chicago, Illinois: University of Chicago Press.

Dalen, H.P.v., & Henkens, K. (2005). Signals in science — on the importance of signaling in gaining attention in science. Scientometrics, 64(2), 209-233.

Davis, M.S. (1971a). That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology. Philosophy of the Social Sciences, 1(2), 309-344

Davis, M.S. (1971b). That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology. Phil. Soc. Sci., 1, 309-344.

Dorigo, M., & Gambardella, L.M. (1997). Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computation, 1(1), 53-66.

Dunbar, K. (1993). Concept discovery in a scientific domain. Cognitive Science, 17, 397-434.

Evans, M., & Kaufman, M. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature, 292(5819), 154-156.

Fleming, L., Mingo, S., & Chen, D. (2007). Collaborative brokerage, generative creativity, and creative success. Administrative Science Quarterly, 52, 443-475.

Fortunato, S., Flammini, A., & Menczer, F. (2006). Scale-free network growth by ranking. Phys. Rev. Lett., 96, 218701.

Freeman, L.C. (1977). A set of measuring centrality based on betweenness. Sociometry, 40, 35-41.

Fuchs, S. (1993). A sociological theory of scientific change. Social Forces, 71(4), 933-953.

Garfield, E. (1992). Of Nobel class: Part 2. Forecasting Nobel Prizes using citation data and the odds against it. Current Contents, 35, 3-12.

Garfield, E., & Welljamsdorof, A. (1992). Of Nobel class — a citation perspective on high-impact research authors. Theoretical Medicine, 13(2), 117-135.

Gill, J. (2005). An entropy measure of uncertainty in vote choice. Electorial Studies, 24, 371-392.

Girvan, M., & Newman, M.E.J. (2002). Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA, 99, 7821-7826.

Gladwell, M. (2007, 01/08/2007). Open secrets: Enron, intelligence, and the perils of too much information. The New Yorker.

Glymour, C. (2004). The automation of discovery. Daedelus, Winter, 69-77.

Goffman, W., & Harmon, G. (1971). Mathematical approach to the prediction of scientific discovery. Nature, 229, 103-104.

Goffman, W., & Newill, V.A. (1964). Generalisation of epidemic theory: An application to the transmission of ideas. Nature, 204, 225-228.

Gordon, M.D., & Lindsay, R.K. (1996). Toward discovery support systems: A replication, re- examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. Journal of the American Society for Information Science, 47(2), 116-128.

Griffith, B.C., & Mullins, N.C. (1977). Coherent social groups in scientific change. Science, 177(4053), 959-964.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. New York, NY.

Guiffre, K. (1999). Sandpiles of opportunity: Success in the art world. Social Forces, 77(3), 815-832.

Hansen, M.T. (1999). The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits. Administrative Science Quarterly, 44(1), 82-111.

Heinze, T., & Bauer, G. (2007). Characterizing creative scientists in nano-S&T: Productivity, multidisciplinarity, and network brokerage in a longitudinal perspective Scientometrics, 70(3), 811-830.

Heinze, T., Shapira, P., Senker, J., & Kuhlmann, S. (2007). Identifying creative research accomplishments: Methodology and results for nanotechnology and human genetics Scientometrics, 70(1), 125-152.

Hilderman, R.J., & Hamilton, H.J. (2001). Knowledge discovery and measures of interest. Norwell, MA: Kluwer Academic Publishers.

Hirsch, J.E. (2005a). An index to quantify an individual's scientific output. PNAS, 102, 16569.

Hirsch, J.E. (2005b). An index to quantify an individual's scientific research output. Proceedings of the national academy of sciences of the United States of America, 102, 16569.

Holton, G., Chang, H., & Jurkowitz, E. (1996). How a scientific discovery is made: A case history. American Scientist, 84(4), 364-375.

Hummon, N.P., & Doreian, P. (1989). Connectivity in a citation network — the development of DNA theory. Social Networks, 11(1), 39-63.

Ioannidis, J.P., & Trikalinos, T.A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. J Clin Epidemiol, 58, 543-549.

Itti, L., & Baldi, P. (2005). A principled approach to detecting surprising events in video. In proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)(pp. 631-637.

Katz, E., & Lazarsfeld, P. (1955). Personal Influence. New York: The Free Press.

Klahr, D., & Simon, H.A. (1999). Studies of scientific discovery: Complementary approaches and convergent findings. Psychological Bulletin, 125(5), 524-543.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 91-101). ACM Press.

Kuhn, T.S. (1962). The structure of scientific revolutions. Chicago: University of Chicago Press.

Kuhn, T.S. (1970). The structure of scientific revolutions. (2nd ed.): University of Chicago Press.

Kullback, S., & Leibler, R.A. (1951). On information and suffciency. Annals of Mathematical Statistics, 22, 79-86.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003). On the bursty evolution of blogspace. In proceedings of the WWW2003(pp. 477). Budapest, Hungary.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2004). Structure and evolution of blogspace. Communications of the ACM, 47(12), 35-39.

Laudan, L., Donovan, A., Laudan, R., Barker, P., Brown, H., Leplin, J., et al. (1986). Scientific change — philosophical models and historical research. Synthese, 69(2), 141-223.

Lazarsfeld, P.F., Berelson, B., & Gaudet, H. (1944). The people's choice: How the voter makes up his mind in a presidential campaign. New York: Columbia University Press.

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. Journal of the American Society for Information Science and Technology, 58(9), 1303-1319.

Liben-Nowell, D., & Kleinberg, J. (2008). Tracing information flow on a global scale using Internet chain-letter data. PNAS, 105(12), 4633-4638.

Lindahal, B.I.B. (1992). Discovery, theory change, and the Nobel Prize: On the mechanism of scientific evolution. Theoretical Medicine, 13(2), 97-231.

Lindsay, R.K., & Gordon, M.D. (1999). Literature-based discovery by lexical statistics. Journal of the American Society for Information Science, 50(7), 574-587.

Liqiang, G., & Howard, J.H. (2006). Interestingness measures for data mining: A survey. ACM Computing Surveys, 38(3), 9.

Lokker, C., McKibbon, K.A., McKinlay, R.J., Wilczynski, N.L., & Haynes, R.B. (2008). Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. BMJ, 336(7645), 655-657.

Marshall, B.J. (2005). Helicobacter connections. Nobel Lecture.

Marshall, B.J., & Warren, J.R. (1984). Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet, 16(1), 1311-1315.

Mayer, R.E. (1995). The search for insight: Grappling with Gestalt Psychology's unanswered questions. In R.J. Sternberg & J.E. Davidson (Eds.), The Nature of Insight (pp. 3-32). Cambridge, MA: The MIT Press.

Morris, S.A., & Van der Veer Martens, B. (2008). Mapping research specialties. Annual Review of Information Science and Technology, 42, 213-295.

Mullins, N.C., Hargens, L.L., Hecht, P.K., & Kick, E.L. (1977). The group structure of cocitation clusters: A comparative study. American Sociological Review, 42(4), 552-562.

Newman, M.E.J. (2001). The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA, 98, 404-409.

Nowakowska, M. (1973). An epidemical spread of scientific objects: an attempt of empirical approach to some problems of meta-science. Theory and Decision, 3, 262-297.

NSF. (2007). Important Notice No. 130: Transformative research. Retrieved Nov 19, 2008, 2008, from http://www.nsf.gov/pubs/2007/in130/in130.jsp

Perkins, D.N. (1995). Insight in minds and genes. In R.J. Sternberg & J.E. Davidson (Eds.), The nature of insight (pp. 495-534). Cambridge, MA: MIT Press.

Perneger, T.V. (2004). Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ. BMJ, 329, 546-547.

Pincock, S. (2005). Nobel Prize winners Robin Warren and Barry Marshall. Lancet, 366(9495), 1429.

Pirolli, P. (2007). Information foraging theory: Adaptive interaction with information. Oxford, England: Oxford University Press.

Radder, H. (1997). Philosophy and history of science: Beyond the Kuhnian paradigm.

Studies in History and Philosophy of Science, 28(4), 633-655.

Redner, S. (2004). Citation statistics from more than a century of Physical Review. Phys. Rev. E (Submitted for Publication).

Sandstrom, P.E. (1999). Scholars as subsistence foragers. Bulletin of the American Society for Information Science, 25(3), 17-20.

Schaffner, K.F. (1992). Theory change in immunology part I: extended theories and scientific progress. Theoretical Medicine, 13(2), 175-189.

Schwarz, J.H. (1982). Superstring theory. Physics Reports-Review Section of Physics Letters, 89(3), 224-322.

Shibata, N., Kajikawa, Y., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles Journal of the American Society for Information Science and Technology, 58(6), 872-882.

Shneiderman, B. (2002). Leonardo's laptop: Human needs and the new computing technologies. Cambridge, MA: MIT Press.

Shneiderman, B. (2007). Creativity support tools: accelerating discovery and innovation. Communications of the ACM, 50(12), 20-32.

Simon, H.A. (1981). The sciences of the artificial. Cambridge, MA: MIT Press.

Simon, H.A., Langley, P.W., & Bradshaw, G.L. (1981a). Scientific discovery as problem-solving. Synthese, 47, 1-27.

Smalheiser, N.R., & Swanson, D.R. (1996). Indomethacin and Alzheimer's disease. Neurology, 46, 583.

Small, H. (1999). A passage through science: Crossing disciplinary boundaries. Library Trends, 48(1), 72-108.

Small, H. (2000). Charting pathways through science: Exploring Garfield's vision of a unified index to science. In B. Cronin & H.B. Atkins (Eds.), The web of knowledge: A festschrift in honor of eugene garfield (pp. 449-473). Medford, NJ: Information Today, Inc.

Small, H., & Crane, D. (1979). Specialties and disciplines in science and social science. Scientometrics, 1, 445-461.

Snijders, T.A.B. (2001). The statistical evaluation of social network dynamics. In M.E. Sobel & M.P. Becker (Eds.), Sociological methodology (pp. 361-395). Boston and London: Basil Blackwell.

Soofi, E.S., & Retzer, J.J. (2002). Information indices: unification and applications. Journal of Econometrics, 107, 17-40.

Stewart, J.A. (1990). Drifting continents and colliding paradigms: Perspectives on the geoscience revolution. Bloomington, IN: Indiana University Press.

Sullivan, D., Koester, D., White, D.H., & Kern, R. (1980). Understanding rapid theoretical change in particle physics: A month-by-month co-citation analysis. Scientometrics, 2(4), 309-319.

Swanson, D.R. (1986a). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine(30), 7-18.

Swanson, D.R. (1986b). Undiscovered public knowledge. Library Quarterly, 56(2), 103-118.

Swanson, D.R. (1987). Two medical literatures that are logically but not bibliographically connected. Journal of the American Society for Information Science, 38, 228-233.

Swanson, D.R. (1988). Migraine and magnesium: Eleven neglected connections. Perspectives in Biology and Medicine, 31, 526-557.

Swanson, D.R., & Smalheiser, N.R. (1999). Implicit text linkages between Medline records: Using Arrowsmith as an aid to scientific discovery. Library Trends, 48, 48-59.

Thagard, P. (1992). Conceptual revolutions. Princeton, New Jersey: Princeton University Press.

Thomas, J.J., & Cook, K.A. (Eds.). (2005). Illuminating the path: The research and development agenda for visual analytics, Los Alamitos, CA: IEEE Computer Society Press.

Valente, T.W. (1996). Social network thresholds in the diffusion of innovations. Social Networks, 18, 69-89.

Wagner-Dobler, R. (1999). William Goffman's "Mathematical approach to the prediction of scientific discovery" and its application to logic, revisited. Scientometrics, 46(3), 635-645.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications. Cambridge University Press.

Chapter 6 Knowledge Domain Analysis

This chapter is concerned with quantitative approaches to the study of emerging trends and changes in science. A key insight is that a domain of knowledge is determined by the perspective we choose to take. This view echoes what we have seen earlier in Chapter 3 about the role of mental models in developing our understanding of the world. We first introduce the principles of progressive knowledge domain analysis. In the second part of the chapter, we describe a new approach called multiple-perspective co-citation analysis. It is designed to shift the traditional focus from the references to papers that have been influenced by such references.

6.1 Progressive Knowledge Domain Visualization

The goal of progressive visualization is to reveal the evolution of an underlying knowledge domain over time. The concept of knowledge domain is defined as the unit of analysis that can adequately represent the essence of the development of the underlying knowledge. Examples of knowledge domains include a topic area of research, a field of study, a discipline, or a combination of any of these entities. The definition is intentionally broad. A knowledge domain cannot exist by itself; its existence depends on our perspectives!

Our mental models determine what may be considered as part of a knowledge domain. A domain of knowledge usually deals with many topics that may appear to be loosely related unless they are seen from a unifying perspective. An existing domain appears to be relatively stable merely because we have used to the same perspective. A new domain may come into being because a creative perspective is found. Such perspectives may be inspired by the external world as well as by our internal world. The notion of a knowledge domain is broader than a paradigm in that a single knowledge domain may accommodate multiple competing paradigms. We use the term *knowledge domain* to emphasize the dynamic nature of the phenomenon. In this chapter, unless stated otherwise, we refer to a network representation of an underlying knowledge domain.

C. Chen, *Turning Points*

6.1.1 Scientific Revolutions

Thomas Kuhn's *Structure of Scientific Revolutions* is widely cited across numerous scientific disciplines. In Kuhn's theory, science evolves by repeatedly going through a series of states, namely, establishing a paradigm, expanding and consolidating the paradigm, the paradigm in crisis, and a revolution — a shift of the paradigm. Kuhn's work has generated deep interests in detecting and tracking evolving and shifting paradigms.

One of the widely studied sources of input for the trails of paradigms is the rich and growing literature of scientific disciplines. Scientific paradigms represent the theories, principles, and methods that dominate the knowledge domain of a scientific field and the community of scientists who work in the field. As Kuhn pointed out, the dynamics of paradigm will be inevitably reflected in the writings of scientists in the field. In addition, the emerging, changing, and competing paradigms will also leave their trails in the literature.

A particularly informative source of clues about such trails in the literature is how scientists reference to earlier work. Since each scientific idea, or contribution, is embodied in the form of a published article in the literature, we can tell a lot about the impact of the original idea from how often and how exactly the article has been cited by peer scientists in subsequent years. The study of citation-related patterns is called citation analysis. Henry Small (1977) studied how research focuses in the field of collagen research changed in terms of how clusters of co-cited references in a network changed over consecutive years. Some of the clusters appeared in one year and disappeared in the next year. Meanwhile, new clusters emerged. Small's study predated many modern visualization techniques, but the language in his description was vivid enough for everyone to picture the changes.

Animated visualization techniques can be used to re-construct citation and co-citation events chronologically so that the history of the development of a scientific field is presented vividly and intuitively (Chen & Kuljis, 2003). The animated visualization enabled us to identify paradigm-like clusters of co-cited articles corresponding to significant changes in the field of superstring, but the visual features of some of the groundbreaking articles were not distinct enough to lend themselves to a simple visual search. Progressive knowledge domain visualization was developed to improve the capability of these techniques so that groundbreaking articles can be characterized by distinguishable visual features (Chen, 2004).

A research front represents the state of the art of a field. Research fronts move along with the underlying scientific field as new articles replace existing articles. The intellectual base consists of all the references cited by a research front. A cluster formed by tightly co-cited references is like a footprint of the research front. As the field moves ahead, a trail of such co-citation clusters, or footprints, can be detected in the literature. Transitions from one paradigm to another are expected to appear as the linkage between clusters corresponding to the underlying paradigms. Figure 6.1 illustrates this scenario.

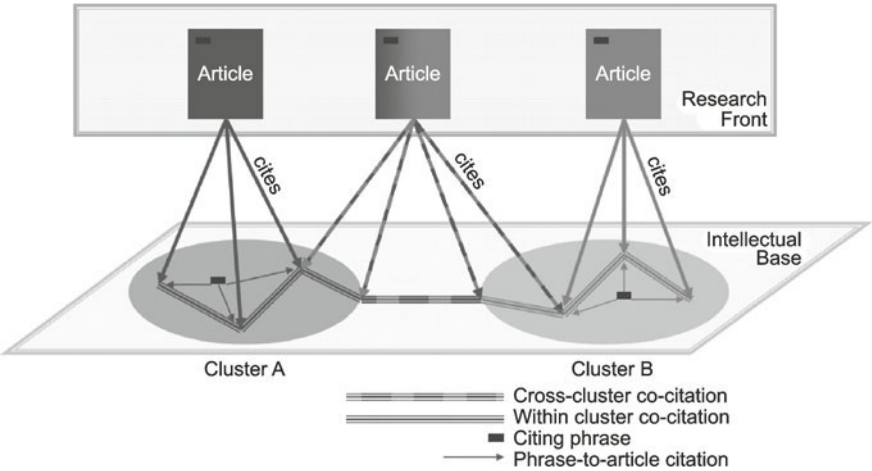


Fig. 6.1 The relationship between a research front and its intellectual base. Source: (Chen, 2006).

6.1.2 Tasks

The knowledge domain visualization has three primary tasks for understanding a body of scientific literature or other types of documents such as grant proposals and patent applications:

1) Improving the clarity of individual networks;
2) Highlighting transitions between adjacent networks;
3) Identifying potentially important nodes.

The first task focuses on the clarity of individual networks' representations. One of the major aesthetic criteria established by research in graph drawing is that link crossings should be avoided whenever possible. A network visualization with the least number of edge crossings is regarded as not only aesthetically pleasing, but also more efficient to work with in terms of the performance of relevant perceptual tasks. The number of link crossings may be reduced by pruning various links in a network. Minimum spanning trees and Pathfinder network scaling are commonly used algorithms. The major advantages and disadvantages of these scaling techniques are further analyzed in the following subsection.

The second task requires that two adjacent networks can be progressively merged so that it becomes clear which part of the earlier network is persistent in the new network, which part of the earlier network is no longer active in the new network, and which part of the new network is completely new. Much of the novelty of our method is associated with how we address this issue.

The third task underlines the role of visually salient features in simplifying search tasks for intellectual turning points. Visually salient nodes include landmark nodes, pivot nodes, and hub nodes.

6.1.2.1 Improving the Clarity of Networks

Co-citation networks often have too many links to show without blocking each other's paths. There are two general approaches to reduce the number of links in a display: a threshold-based approach and a topology-based approach. In a threshold-based approach, the elimination of a link is purely determined by whether the link's weight exceeds a threshold. In contrast, in a topology-based approach, the elimination of a link is determined by a more extensive consideration of intrinsic topological properties; therefore, such approaches tend to preserve certain topological intrinsic properties more reliably, although the computational complexity tends to be higher.

Pathfinder network scaling is originally developed by cognitive scientists to build procedural models based on subjective ratings (Schvaneveldt, 1990). It uses a more sophisticated link elimination mechanism than a minimum spanning tree (MST) algorithm. It retains the most important links and preserves the integrity of the network. Every network has a unique Pathfinder network, which contains all the alternative MSTs of the original network.

Pathfinder network scaling aims to simplify a dense network while preserving its salient properties. The topology of a Pathfinder network is determined by two parameters r and q. The r parameter defines a metric space over a given network based on the Minkowski distance so that one can measure the length of a path connecting two nodes in the network. The Minkowski distance becomes the familiar Euclidean distance when $r = 2$. When $r = \infty$, the weight of a path is defined as the maximum weight of its component links, and the distance is known as the maximum value distance. Given a metric space, a triangle inequality can be defined as follows,

$$w_{ij} \leqslant (\Sigma_k w^r_{n_k n_{k+1}})^{1/r}$$

where w_{ij} is the weight of a direct path between i and j, $w_{n_k n_{k+1}}$ is the weight of a path between n_k and n_{k+1}, for $k = 1, 2, \ldots, m$. In particular, $i = n_1$ and $j = n_k$. In other words, the alternative path between i and j may go all the way round through nodes $n_1, n_2, \ldots, n_k$ as long as each intermediate links belong to the network.

If w_{ij} is greater than the weight of alternative path, then the direct path between i and j violates the inequality condition. Consequently, the link $i - j$ will be removed because it is assumed that such links do not represent the most salient aspects of the association between the nodes i and j.

The q parameter specifies the maximum number of links that alternative paths can have for the triangle inequality test. The value of q can be set to any integer between 2 and $N - 1$, where N is the number of nodes in the network. If an alternative path has a lower cost than the direct path, the direct path will be removed. In this way, Pathfinder reduces the number of links from the original network, while all the nodes remain untouched. The resultant network is also known as a minimum-cost network.

The strength of Pathfinder network scaling is its ability to derive more accurate local structures than other comparable algorithms such as multidimensional scaling (MDS) and minimum spanning tree (MST). However, the Pathfinder algorithm is computationally expensive. The maximum pruning power of Pathfinder is achievable with $q = N-1$ and $r = \infty$; not surprisingly, this is also the most expensive one because all the possible paths must be examined for each link. Some recent implementations of Pathfinder networks reported the use of the set union of MSTs.

6.1.2.2 Merging Heterogeneous Networks

Intellectual structures of a knowledge domain before and after a major conceptual revolution can be fundamentally different as new theories and new evidence become predominant. Co-citation networks of citation classics in a field may differ from co-citation networks of newly published articles. The key question is: what is the most informative way to merge potentially diverse networks?

A merged network needs to capture the important changes over time in a knowledge domain's co-citation structure. We need to find when and where the most influential changes take place so that the evolution of the domain can be characterized and visualized. Few studies in the literature investigated network merge from a domain-centric perspective. The central idea of our method is to visualize how different network representations of an underlying phenomenon can be informatively stitched together.

6.1.2.3 Visually Salient Nodes in Merged Networks

The importance of a node in a co-citation network can be quickly identified by the local topological structure of the node and by additional attributes of the node. We are particularly interested in three types of nodes: 1) landmark nodes, 2) hub nodes, and 3) pivot nodes (see Fig. 6.2).

A landmark node is a node that has extraordinary attribute values. For example, a highly cited article tends to provide an important landmark regardless how it is co-cited with other articles. Landmark nodes can be rendered by distinctive visual-spatial attributes such as size, height, or volume. A hub node has a relatively large node degree. A widely co-cited article is a good candidate for significant intellectual contributions. A high-degree hub-like node is also easy to recognize in a visualized network. Both landmark nodes and hub nodes are commonly used in network visualization. Although the concept of pivot nodes is available in various contexts, the way they are used in our method is novel. Pivot nodes are joints between different networks. They are either the common nodes shared by two networks, or the gateway nodes that are connected by inter-network links. Pivot nodes have an essential role in our method.

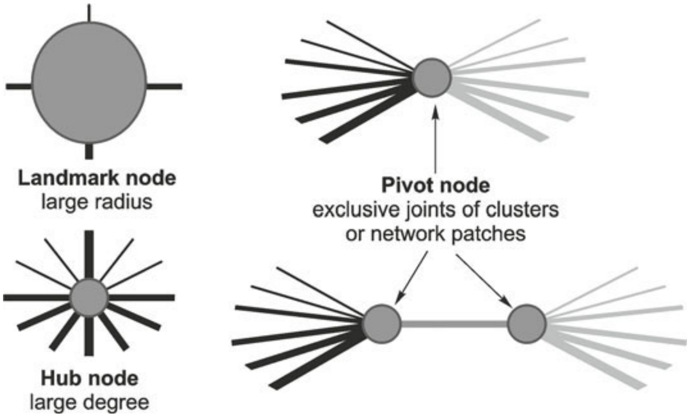


Fig. 6.2 Three types of salient nodes in a co-citation network. Source: (Chen, 2004).

6.1.3 CiteSpace[1]

CiteSpace has been the primary vehicle for progressive knowledge domain analysis. It is a freely available Java application for visualizing and analyzing emerging patterns and critical changes in the literature of a scientific domain (Chen, 2004; Chen, 2006; Chen, Ibekwe-SanJuan, & Hou, 2010). CiteSpace uses a set of bibliographic records as input, typically including information on cited references, and produces interactive visualizations of networks of authors, references, and several other types of entities as types of nodes over a number of consecutive time slices. These visualizations are designed to help users identify intellectual turning points, critical paths of transitions, and aggregations of individual nodes. The general procedure is described below (see Fig. 6.3). Details of more specific analytic features will be given as they are needed.

6.1.3.1 Time Slicing

Time slicing divides the entire time interval into equal-length segments called time slices. The duration of each segment can be as short as one year or as long as tens and even hundreds of years. If appropriate data is available, it is possible to slice it thinner to make monthly or weekly segments. Currently, sliced segments are mutually exclusive, although overlapping segments could be an interesting alternative.

6.1.3.2 Sampling

Citation analysis and co-citation analysis typically sample the most highly cited work — the cream of crop. In order to construct a network in CiteS-

[1] http://cluster.ischool.drexel.edu/∼cchen/citespace

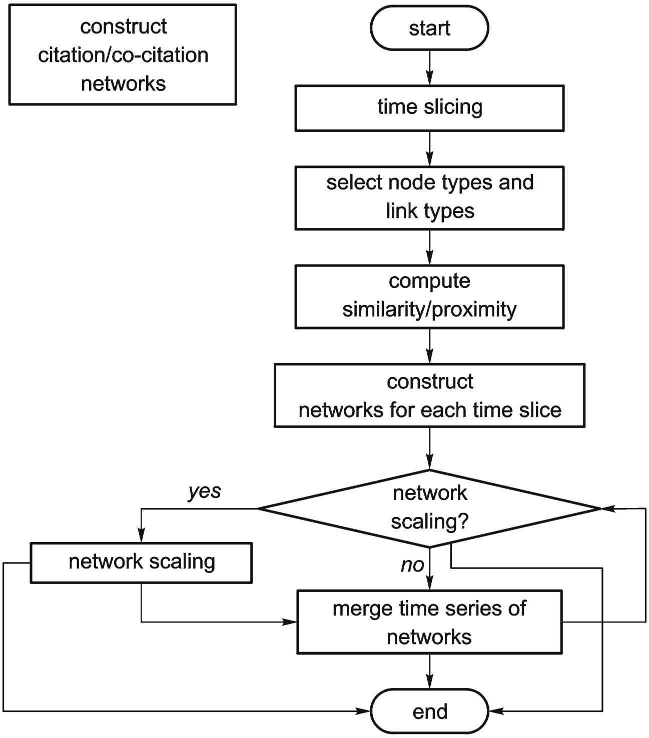

construct citation/co-citation networks
start
time slicing
select node types and link types
compute similarity/proximity
construct networks for each time slice
network scaling?
yes
no
network scaling
merge time series of networks
end

Fig. 6.3 The general procedure supported in CiteSpace.

pace, users may set their own criteria for node selection and link selection. Alternatively, they can use the default setting provided by CiteSpace. The simplest way to select nodes is the *Top-N* method, in which the *N* most cited articles within the timeframe of each slice will be included in the final network. Similarly, the *Top-N%* method will include the *N%* of the most cited references within each slice. CiteSpace also allows the user to choose three sets of threshold values and interpolates these values across all the slices. Each set of threshold values are: a citation count (c), a co-citation count (cc), and a cosine coefficient of co-citation similarity (ccv). In CiteSpace, the user needs to select desired thresholds in the beginning, the middle, and the ending slices. CiteSpace automatically assigns interpolated thresholds to the remaining slices.

Research has shown that citation counts often follow a power law distribution. The vast majority of published articles are never cited. On the other hand, a small number of articles dominate a lion share of citations. Many factors may influence the frequency and distribution of citations to published articles. A highly cited article is highly visible. Its visibility is likely to attract more citations. As far as intellectual turning points are concerned, we are particularly interested in articles that have rapidly growing citations. In the following superstring example, we use a simple model to normalize the citations of an article within each time slice by the logarithm of its publication age — the number of years elapsed since its publication year. The rationale is to highlight articles that increased most in the early years of their publication.

6.1.3.3 Modeling

By default, co-citation counts are calculated within each time slice. Cocitation counts are normalized as cosine coefficients, provided $c(i) \neq 0$ and $c(j) \neq 0$:

$$cc_{cosine}(i, j) = \frac{cc(i, j)}{\sqrt{c(i) * c(j)}}$$

where $cc(i, j)$ is the co-citation count between documents i and j, and $c(i)$ and $c(j)$ are their citation counts, respectively. The user can specify a selection threshold for co-citation coefficients; the default value is 0.15.

Alternative measures of co-citation strengths are available in the information science literature, such as Dice and Jaccard coefficients.

6.1.3.4 Pruning

An effective pruning can reduce link crossing and improve the clarity of the resultant network visualization. CiteSpace supports two common network pruning algorithms, namely Pathfinder and MST. The user can select to prune individual networks only, or the merged network only, or both. Pruning increases the complexity of the visualization process. In the following section, visualizations with local pruning and global pruning are discussed.

Here we concentrate on Pathfinder-based pruning. To prune individual networks with Pathfinder, the parameters q and r are set to $N_k - 1$ and ∞, respectively, to ensure the most extensive pruning effect, where N_k is the size of the network in the kth time slice. For the merged network, the q parameter is $(\Sigma N_k) - 1$, for $k = 1, 2, \ldots$.

6.1.3.5 Merging

The sequence of time sliced networks is merged into a synthesized network, which contains the set union of all nodes ever appear in any of the individual networks. Links from individual networks are merged based on either the earliest establishment rule or the latest reinforcement rule. The earliest establishment rule selects the link that has the earliest time stamp and drops subsequent links connecting the same pair of nodes, whereas the latest reinforcement rule retains the link that has the latest time stamp and eliminates earlier links.

By default, the earliest establishment rule applies. The rationale is to support the detection of the earliest moment when a connection was made in the literature. More precisely, such links mark the first time a connection becomes strong enough with respect to the chosen thresholds.

6.1.3.6 Mapping

The layout of each network, either individual time sliced networks or the merged one, is produced using Kamada and Kawa's algorithm (Kamada & Kawai, 1989). The size of a node is proportional to the normalized citation counts in the latest time interval. Landmark nodes can be identified by their large discs. The label size of each node is proportional to citations of the article, thus larger nodes also have larger-sized labels. The user can enlarge font sizes at will. The width of a link is proportional to the corresponding co-citation coefficient. The color of a link indicates the earliest appearance time of the link with reference to chosen thresholds.

Visually salient nodes such as landmarks, hubs, and pivots are easy to detect by visual inspection. CiteSpace also includes algorithms to detect such nodes computationally. The visual effect is a natural result of slicing and merging, while additional computational metrics enhance the visual features even further. A useful computational metric should reflect the degree of a node, and it should also take into account the heterogeneity of the node's links. The more dissimilar links a node connects to others, the more likely the node has a pivotal role to play.

6.1.3.7 Case Study: Superstring

Two superstring revolutions are documented over the last two decades: one in mid-1980s and one in mid-1990s (Schwarz, 1982; Schwarz, 1996). The superstring dataset includes citation data between 1985 and 2003. We asked the leading scientists in the field of superstring to validate visualized networks. We showed the merged map, without pruning, to John Schwarz at CalTech and Edward Witten at Princeton University. Schwarz was the co-author of the article that triggered the first superstring revolution. Witten has a number of highly cited articles on superstring. He is also the top ranked physicist in a list of the 1,000 most cited physicists between 1981 and 1997. The list was compiled by the Institute for Scientific Information (ISI). They were asked to explain the nature of intellectual contributions identified by pivot points and hubs in the networks.

The 19-year time interval was sliced into 6 three-year segments, starting from 1985 – 1987 and ending at 2000 – 2002, plus a one-year segment for 2003. Two sets of results were generated from two separate runs. One used relatively higher threshold settings, which resulted in small networks (Fig. 6.4). The other used lower threshold settings for larger networks (Fig. 6.5). Links were color-coded by the earliest establishment rule. Darker colors indicate links from earlier time slices, whereas lighter colors indicate links from more recent slices.

As shown in Fig. 6.4, the 1984 Green-Schwarz article is a typical pivot node — it is the only contact point between two densely connected clusters in blue (1985 – 1987). It was this article that triggered the first superstring revolution — the famous 1984 Green-Schwarz anomaly cancellation paper.

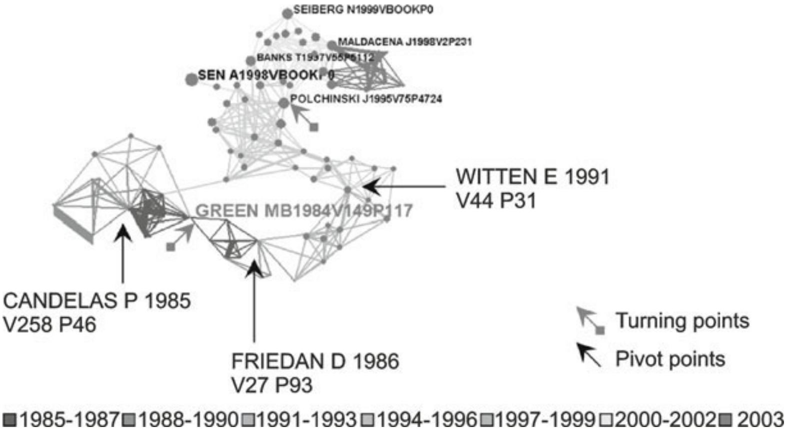


Fig. 6.4 Turning points in superstring research. Source: (Chen, 2004). (see color figure at the end of this book)

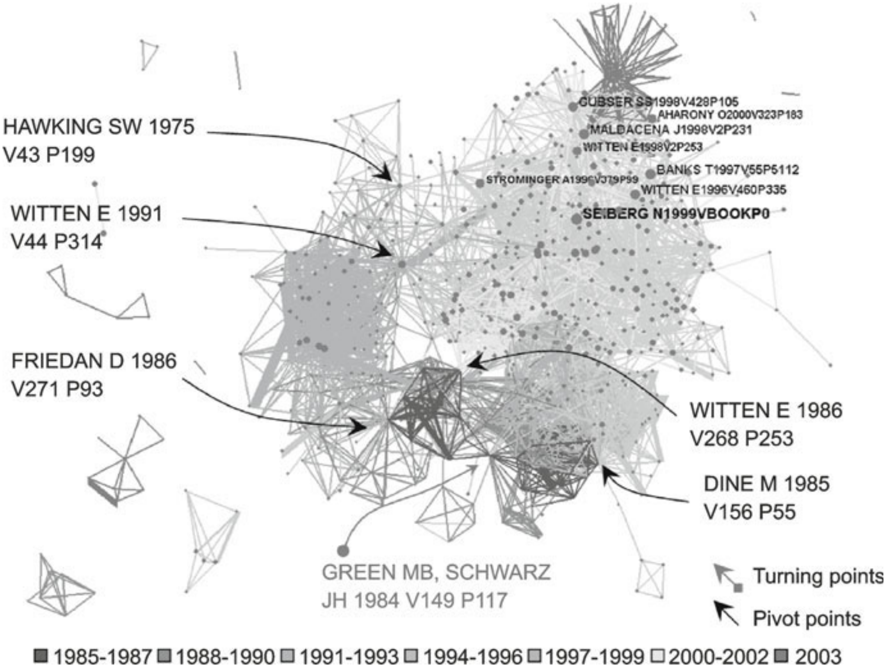


Fig. 6.5 A network of 624 co-cited references. Source: (Chen, 2004). (see color figure at the end of this book)

Friedan's 1986 article is a distinct pivot node connecting a blue cluster (1985 – 1987), a pink cluster (1988 – 1990), and a green cluster (1991 – 1993). Witten's 1986 article is a pivot between a blue cluster (1985 – 1987) and a yellow cluster (2000 – 2002).

In Fig. 6.4, small clusters in red (2003) indicate the candidates for emerging clusters. We were able to find Polchinski's 1995 article in a smaller sized merged network, but the article was overwhelmed by the 4,000 strong links of the larger network. Nevertheless, the quality of the visualized network is promising: intellectually significant articles tend to have topologically unique positions.

Articles by Maldacena, Witten, and Gubser-Klebanov-Polyako, located towards the top of the major network component, were all published in 1998. When we asked Witten to comment an earlier version of the map, in which citation counts were not normalized by years since publication, he indicated that the Green-Scharwz article is more important to the field than the three top cited ones, and that the earlier articles in the 1990s appeared to be underrepresented in the map. There is an apparent mismatch between citation frequencies of nodes and their importance judged by domain experts. Witten's comments raised an important question: is it possible that an intellectually significant article may not always be the most highly cited? Yes, indeed; it is possible.

The comments from domain experts have confirmed that both versions of the merged network indeed highlight significant articles. And these articles tend to have unique topological properties that distinguish themselves from other articles.

6.1.3.8 Other Examples

We have conducted a series of other case studies using CiteSpace, including mass extinctions, terrorism research (Chen, 2006), Sloan Digital Sky Survey (SDSS) in astronomy (Chen, Zhang, & Vogeley, 2009b), and information science (Chen et al., 2010). Here we highlight some of the most representative ones.

Fig. 6.6 shows a visualization of a synthesized network of co-cited references in terrorism research. The overall network is dominated by three tight clusters. Each corresponds to a paradigm. The one at the bottom was formed after the terrorist attacks on September 11, 2001. The one on the left was formed much earlier, which is primarily on physical injuries resulted from terrorist attacks in early 1990s. The cluster on the right was formed by articles on the preparedness of health care concerning the threats of biological and chemical weapons. The transition from the physical injury cluster to the preparedness cluster was characterized by a single article labeled as the turning point in the visualization. It is worth noting that the view of the transition linkage between the two clusters might not exist at the level of individual researchers if no one has ever made a citation chain that contains at least one article in each of the clusters and the turning point article. It is quite possible that researchers working on each cluster do not know any articles in the other cluster, except the common turning point article. In other words, synthesized visualizations like this can reveal something that no individual would be able to see otherwise.

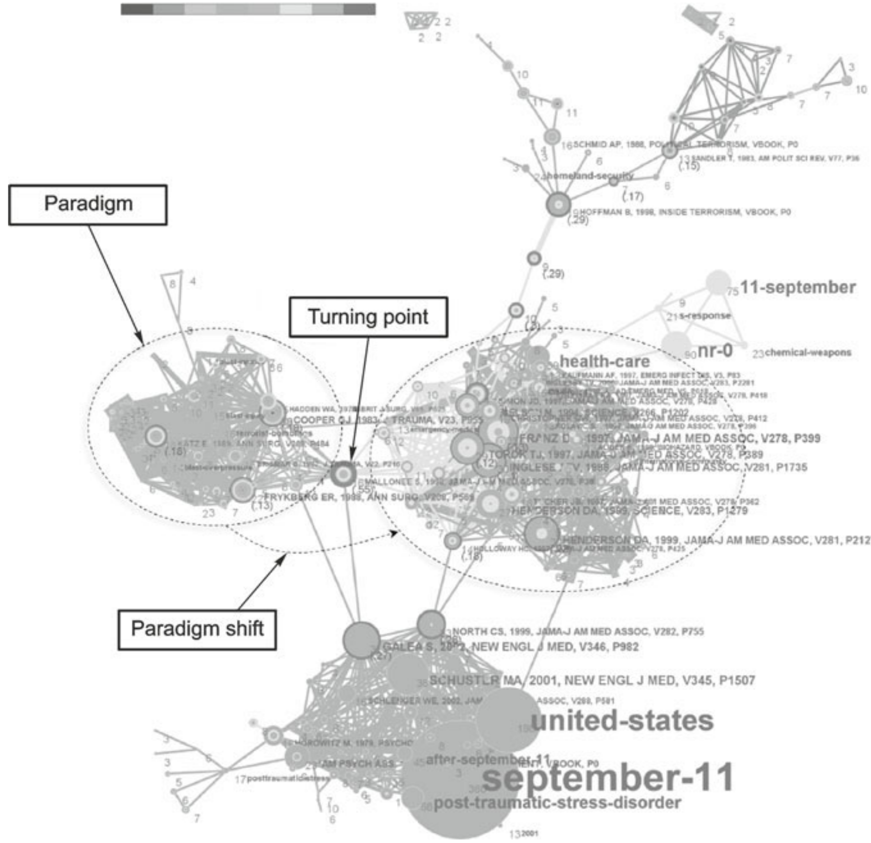

Paradigm
Turning point
Paradigm shift
SCHMID AP, 1988, POLITICAL TERRORISM, VBOOK, P0
SANDLER T, 1983, AM POLIT SCI REV, V77, P36
homeland-security
HOFFMAN B, 1998, INSIDE TERRORISM, VBOOK, P0
11-september
s-response
nr-0
chemical-weapons
health-care
KAUFMANN AF, 1997, EMERG INFECT DIS, V3, P83
FRANZ D, 1997, JAMA-J AM MED ASSOC, V278, P399
TOROK TJ, 1997, JAMA-J AM MED ASSOC, V278, P389
INGLESBY TV, 1999, JAMA-J AM MED ASSOC, V281, P1735
HENDERSON DA, 1999, SCIENCE, V283, P1279
HENDERSON DA, 1999, JAMA-J AM MED ASSOC, V281, P2127
COOPER GJ, 1983, J TRAUMA, V23, P955
FRYKBERG ER, 1988, ANN SURG, V208, P569
MALLONEE S, 1996, JAMA-J AM MED ASSOC
NORTH CS, 1999, JAMA-J AM MED ASSOC, V282, P755
GALEA S, 2002, NEW ENGL J MED, V346, P982
SCHUSTER MA, 2001, NEW ENGL J MED, V345, P1507
united-states
after-september-11
september-11
post-traumatic-stress-disorder
posttraumaticstress
2001

Fig. 6.6 Major areas in terrorism research. Source: (Chen, 2006). (see color figure at the end of this book)

Another interesting case study is a visualization of research concerning mass extinctions. A historical view of the research area is shown in Fig. 6.7. The visualization depicts two major lines of research. The one that is labeled as KTB(65Ma) started from the far left and comes to an end around 1993. The other line, labeled as PTB(250Ma), emerged after 1991. It appears that the research community shifted its focus from the KTB thread to the PTB thread. Our case study revealed that the two threads of research shared a considerable degree of similarity. Both threads started with a theory and the mission was to search for diagnostic evidence. The PTB thread was clearly inspired by the success of the KTB thread. We reported this macroscopic pattern in an article published in 2006. Intriguingly, as shown in Fig. 6.8, an article published in 2010 by mass extinction experts included an almost identical summary of the transition (French & Koeberl, 2010). This is particularly encouraging given that we were able to identify the same pattern purely based on progressive knowledge domain analysis and with no expertise in the subject.

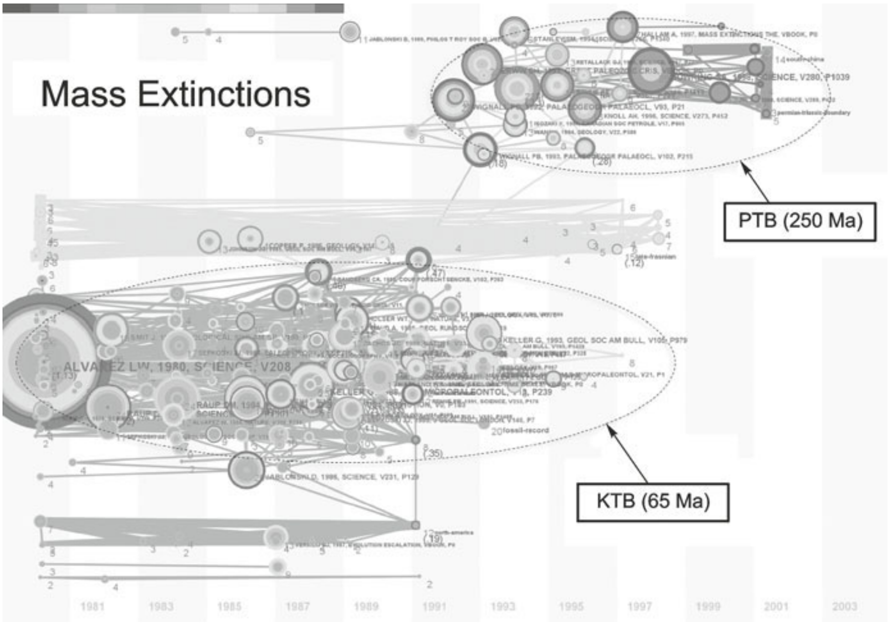
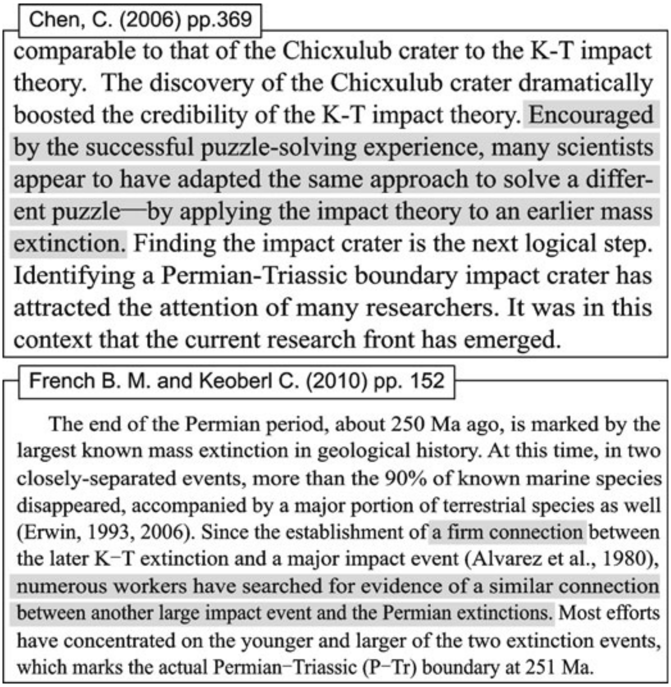


Fig. 6.7 Trends in mass extinctions research. Source: (Chen, 2006). (see color figure at the end of this book)



Chen, C. (2006) pp.369

comparable to that of the Chicxulub crater to the K-T impact theory. The discovery of the Chicxulub crater dramatically boosted the credibility of the K-T impact theory. Encouraged by the successful puzzle-solving experience, many scientists appear to have adapted the same approach to solve a different puzzle—by applying the impact theory to an earlier mass extinction. Finding the impact crater is the next logical step. Identifying a Permian-Triassic boundary impact crater has attracted the attention of many researchers. It was in this context that the current research front has emerged.

French B. M. and Keoberl C. (2010) pp. 152

The end of the Permian period, about 250 Ma ago, is marked by the largest known mass extinction in geological history. At this time, in two closely-separated events, more than the 90% of known marine species disappeared, accompanied by a major portion of terrestrial species as well (Erwin, 1993, 2006). Since the establishment of a firm connection between the later K−T extinction and a major impact event (Alvarez et al., 1980), numerous workers have searched for evidence of a similar connection between another large impact event and the Permian extinctions. Most efforts have concentrated on the younger and larger of the two extinction events, which marks the actual Permian−Triassic (P−Tr) boundary at 251 Ma.

Fig. 6.8 Macroscopic patterns were identified by our citation analysis published in 2006 and by mass extinction experts in 2010.

6.2 A Multiple-Perspective Co-Citation Analysis

A multiple-perspective co-citation analysis addresses an important issue that has been overlooked by the traditional co-citation analysis. The method consists of the analysis of structural, temporal, and semantic patterns as well as the use of both citing and cited items for interpreting the nature of co-citation clusters. We illustrate the principles of the new approach with an author co-citation and a document co-citation as the basis. We first set the context of our work with reference to the traditional procedure and introduce several citation-related and structure-related metrics for subsequent discussions. Then we explain three components of the new procedure, namely, clustering, labeling, and sentence selection.

6.2.1 Extending the Traditional Procedure

We often take for granted that we can always tell the source of a shadow by just looking at the shadow alone. Henry Bursill's book *Hand Shadows to be Thrown Upon the Wall* is full of vivid shadows made by bare hands on the wall. Fig. 6.9 shows one of the shadows from the book. Making a vividly looking shadow out of something drastically different has become an art. As shown in Fig. 6.10, the motorcycle-shaped shadow is not a shadow of a motorcycle; instead, the source of the shadow was a pile of chunk. Even our natural language becomes awkward to express the split between a shadow and its source. This type of projection is so vividly and precisely rendered that it is hard for us to realize that the boy or the motorcycle does not even exist!

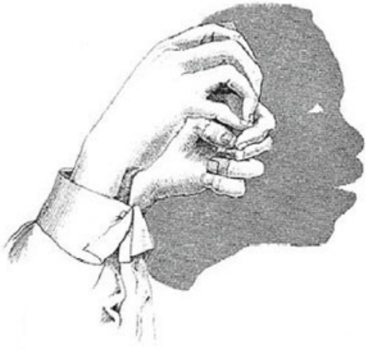


Fig. 6.9 A hand shadow of a boy. Source: (Bursill, 1859).

In an abstract world, the relationship between a shadow and its source can be even more subtle. The traditional approach to the study of co-citation

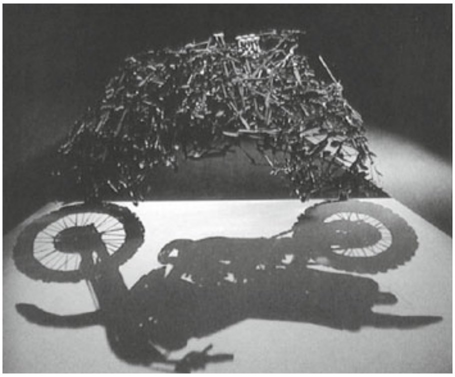


Fig. 6.10 Is this the shadow of a motorcycle.[2]

networks focuses on one of the many possible shadows of a research front rather than the research front itself. Each shadow is formed by a particular perspective, which in turn offers a framework for interpreting the meaning of the research front. But it is not always valid to trace from a shadow back to its source, especially in situations where we need to interpret how a shadow was formed. Using a multiple-perspective approach, we intend to shift the focus from the shadow to the figure that throws the shadow as well as the shadow itself.

The primary goal of co-citation analysis is to identify the intellectual structure of a scientific knowledge domain in terms of the groupings formed by accumulated co-citation trails in scientific literature. The traditional procedure of co-citation analysis for both document co-citation analysis (DCA) and author co-citation analysis (ACA) consists of the following steps:

1) Retrieve citation data from sources such as the *Science Citation Index* (SCI), *Social Science Citation Index* (SSCI), *Scopus*, and *Google Scholar*.
2) Construct a matrix of co-cited references (DCA) or authors (ACA).
3) Represent the co-citation matrix as a node-and-link graph or as a multidimensional scaling (MDS) configuration with possible link pruning using Pathfinder network scaling or minimum spanning tree algorithms.
4) Identify specialties in terms of co-citation clusters, multivariate factors, principle components, or dimensions of a latent semantic space using a variety of algorithms for clustering, community-finding, factor analysis, principle component analysis, or latent semantic indexing.
5) Interpret the nature of co-citation clusters.

The interpretation step is the weakest link. It is time-consuming and cognitively demanding, requiring a substantial level of domain knowledge and synthesizing skills. In addition, much of attention routinely focuses on co-citation clusters per se, but the role of citing articles that are responsible for the formation of such co-citation clusters may not be always investigated

[2] http://epicr.com/wp-content/uploads/2010/07/motorcycle-sculpture_sm1.jpg

as an integral part of a specialty. While the focus on the target of citation may reveal seminal members of a specialty, it does not necessarily reflect the dynamics of the specialty in terms of the impact on its citers. Researchers have used citing information to summarize the essence of a co-citation cluster. For instance, Small (1986) introduced a method for generating specialty narratives by walking through a co-citation network and selecting passages that cite the core documents in a co-citation cluster. There are also other studies that take citing information into account (Schneider, 2009).

Given the diversity and complexity of relationships between citers and cited entities (Cronin, 1981), synthesizing the nature of a co-citation cluster is cognitively too demanding for analysts to handle manually. The lack of algorithmic support for these tasks forces analysts to rely on their own domain knowledge and their experience. It makes it hard to differentiate evidence-based findings from heuristics and speculations. Such ambiguity may hinder subsequent evaluation and scholarly communication of research findings. These problems motivate us to develop a multiple-perspective method to improve the robustness of the traditional procedure.

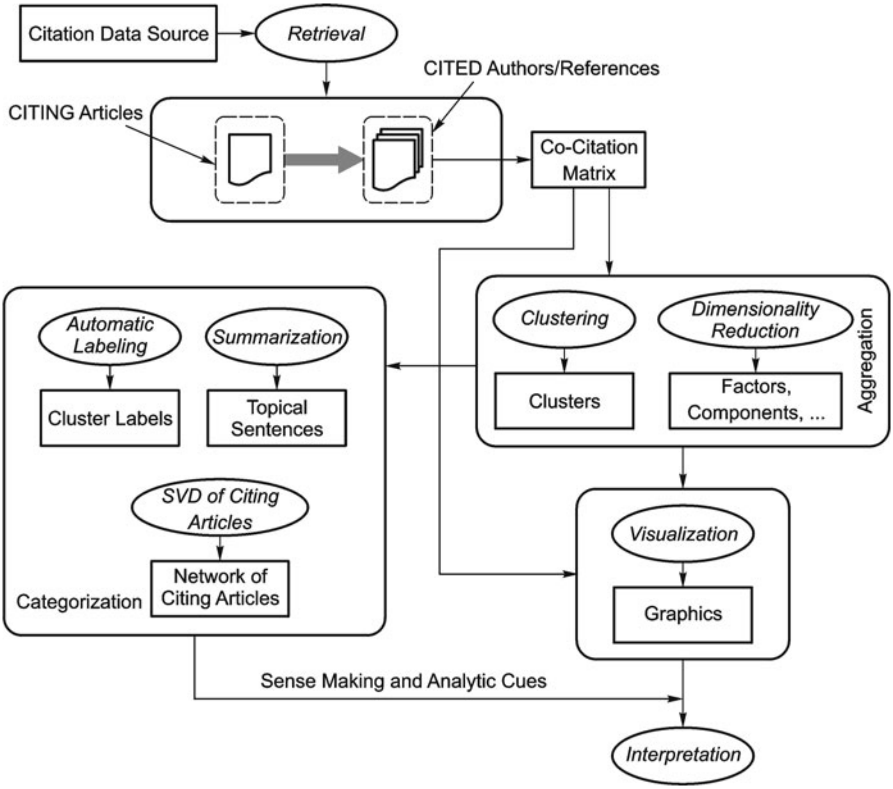


Fig. 6.11 The procedure of a multiple-perspective analysis. Source: (Chen et al., 2010).

The multiple-perspective approach extends and enhances traditional co-citation methods in two ways: (1) by integrating structural and content analysis components sequentially into the new procedure and (2) by facilitating analytic tasks and interpretation with automatic cluster labeling and summarization functions. The key components of the new procedure are highlighted in Fig. 6.11, including clustering, automatic labeling, summarization, and latent semantic models of the citing space (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990).

6.2.2 Metrics

Our new procedure adopts several structural and temporal metrics of co-citation networks and subsequently generated clusters. Structural metrics include betweenness centrality, modularity, and silhouette. Temporal and hybrid metrics include citation burstness and novelty.

The betweenness centrality metric is defined for each node in a network. It measure the extent to which the node is in the middle of a path that connects other nodes in the network (Brandes, 2001; Freeman, 1977). High betweenness centrality values identify potentially revolutionary scientific publications (Chen, 2005) as well as gatekeepers in social networks. If a node provides the only connection between two large but otherwise unrelated clusters, then this node would have a very high value of betweenness centrality. Recently, power centrality introduced by Bonacich (1987) is also drawing a lot of attention in dealing with networks in which someone's power depends on the power of those he/she is socially related to, for example, in (Kiss & Bichler, 2008).

The modularity Q measures the extent to which a network can be divided into independent blocks, i.e. modules (Newman, 2006; Shibata, Kajikawa, Taked, & Matsushima, 2008). The modularity score ranges from 0 to 1. A low modularity suggests a network that cannot be reduced to clusters with clear boundaries, whereas a high modularity may imply a well-structured network. On the other hand, networks with modularity scores of 1 or very close to 1 may turn out to be some trivial special cases where individual components are simply isolated from one another. Since the modularity is defined for any network, one may compare different networks in terms of their modularity, for example, between ACA and DCA networks.

The silhouette metric (Rousseeuw, 1987) is useful in estimating the uncertainty involved in identifying the nature of a cluster. The silhouette value of a cluster, ranging from -1 to 1, indicates the uncertainty that one needs to take into account when interpreting the nature of the cluster. The value of 1 represents a perfect separation from other clusters. In this study, we expect that cluster labeling or other aggregation tasks will become more straightforward for clusters with the silhouette value in the range of 0.7~0.9 or higher.

Burst detection determines whether a given frequency function has statistically significant fluctuations during a short time interval within the overall time period. It is valuable for citation analysts to detect whether and when the citation count of a particular reference has surged. For example, after the 911 terrorist attacks, citations to earlier studies of Oklahoma City Bombing were increased abruptly (Chen, 2006). It can be also used to detect whether a particular connection has been significantly strengthened within a short period of time (Kumar, Novak, Raghavan, & Tomkins, 2003). We adopt the burst detection algorithm introduced in (Kleinberg, 2002).

Sigma (Σ) is introduced in (Chen, Chen, Horowitz, Hou, Liu, & Pellegrino, 2009a) as a measure of scientific novelty. It identifies scientific publications that are likely to represent novel ideas according to two criteria of transformative discovery. As demonstrated in case studies (Chen et al., 2009a), Nobel Prize and other award winning research tends to have highest values of this measure. CiteSpace currently uses $(centrality + 1)^{burstness}$ as the Σ value so that the brokerage mechanism plays more prominent role than the rate of recognition by peers.

6.2.3 Clustering

We adopt a hard clustering approach such that a co-citation network is partitioned to a number of non-overlapping clusters. It is more efficient to use non-overlapping clusters than overlapping ones to differentiate the nature of different co-citation clusters, although it is conceivable to derive a soft clustering version of this particular component. Resultant clusters are subsequently labeled and summarized.

Co-citation similarities between items i and j are measured in terms of cosine coefficients. If A is the set of papers that cites i and B is the set of papers that cite j, then $w_{ij} = \dfrac{|A \bigcap B|}{\sqrt{|A| \times |B|}}$, where $|A|$ and $|B|$ are the citation counts of i and j, respectively; and $|A \bigcap B|$ is the co-citation count, i.e. the number of times they are cited together. Alternative similarity measures are also available. For example, Small (1973) used $w_{ij} = \dfrac{|A \bigcap B|}{|A \bigcup B|}$, which is known as the Jaccard index (Jaccard, 1901).

A good partition of a network would group strongly connected nodes together and assign loosely connected ones to different clusters. This idea can be formulated as an optimization problem in terms of a cut function defined over a partition of a network. Technical details are given in relevant literature (Luxburg, 2006; Ng, Jordan, & Weiss, 2002; Shi & Malik, 2000). A *partition* of a network G is defined by a set of sub-graphs $\{G_k\}$ such that

$G = \bigcup_{k=1}^{K} G_k$ and $G_i \bigcap G_j = \varnothing$, for all $i \neq j$. Given sub-graphs A and B, a *cut function* is defined as follows: $cut(A, B) = \sum_{i \in A, j \in B} w_{ij}$, where w_{ij}'s are the cosine coefficients mentioned above. The criterion that items in the same cluster should have strong connections can be optimized by maximizing $\sum_{k=1}^{K} cut(G_k, G_k)$. The criterion that items between different clusters should be only weakly connected can be optimized by minimizing $\sum_{k=1}^{K} cut(G_k, G - G_k)$. In this study, the cut function is normalized by $\sum_{k=1}^{K} \frac{cut(G_k, G - G_k)}{vol(G_k)}$ to achieve more balanced partitions, where $vol(G_k)$ is the sum of the weights of links in G_k, i.e. $vol(G_k) = \sum_{i \in G_k} \sum_{j} w_{ij}$ (Shi & Malik, 2000).

Spectral clustering is an efficient and generic clustering method (Luxburg, 2006; Ng et al., 2002; Shi & Malik, 2000). It has roots in spectral graph theory. Spectral clustering algorithms identify clusters based on eigenvectors of Laplacian matrices derived from the original network. Spectral clustering has several desirable features compared to traditional algorithms such as k-means and single linkage (Luxburg, 2006):

1) It is more flexible and robust because it does not make any assumptions on the forms of the clusters;
2) It makes use of standard linear algebra methods to solve clustering problems;
3) It is often more efficient than traditional clustering algorithms.

The multiple-perspective method utilizes the same spectral clustering algorithm for both ACA and DCA studies. Despite its limitations (Luxburg, Bousquet, & Belkin, 2009), spectral clustering provides clearly defined information for subsequent automatic labeling and summarization to work with. In this study, instead of letting the analyst to specify how many clusters there should be, the number of clusters is uniformly determined by the spectral clustering algorithm based on the optimal cut described above.

6.2.4 Automatic Cluster Labeling

Candidates of cluster labels are selected from noun phrases and index terms of citing articles of each cluster. These terms are ranked by three different algorithms. In particular, noun phrases are extracted from titles and abstracts of citing articles. The three term ranking algorithms are *tf*idf* (Salton, Yang,

& Wong, 1975), *log-likelihood ratio* (LLR) tests (Dunning, 1993), and *mutual information* (MI). Labels selected by *tf*idf* weighting tend to represent the most salient aspect of a cluster, whereas those chosen by log-likelihood ratio tests and mutual information tend to reflect a unique aspect of a cluster.

Garfield (1979) has discussed various challenges of computationally selecting the most meaningful terms from scientific publications for subject indexing. Indeed, the notion of citation indexing was originally proposed as an alternative strategy to deal with some of the challenges. White (2007a, 2007b) offers a new way to capture the relevance of a communication in terms of the widely known *tf*idf* formula.

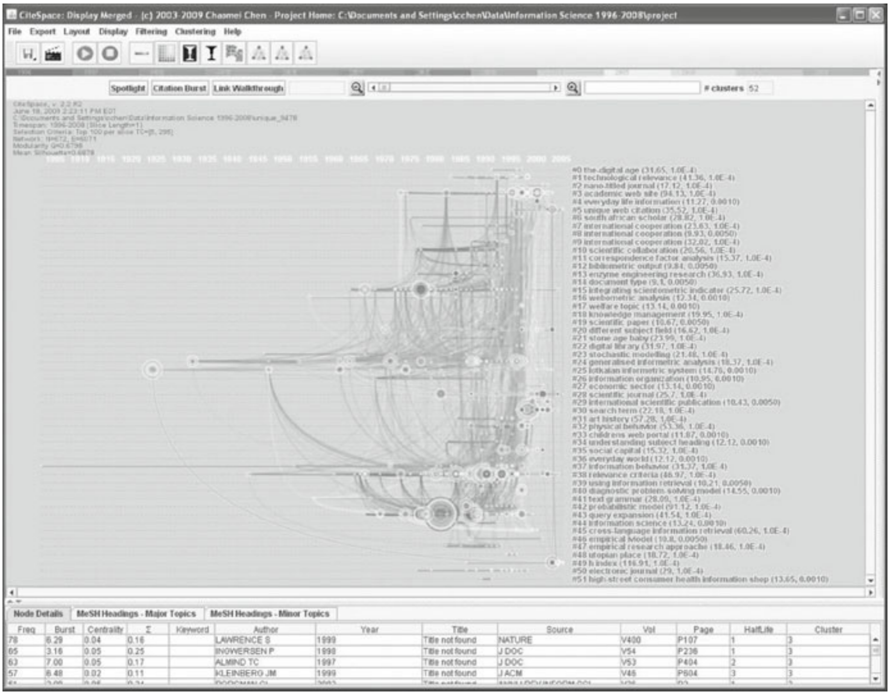
A good text summary should have a sufficient and balanced coverage with minimal redundant information (Sparck Jones, 1999). Teufel and Moens (2002) proposed an intriguing strategy for summarizing scientific articles based on the rhetorical status of statements in an article. Their strategy specifically focuses on identifying the new contribution of a source article and its connections to earlier work. Automatic summarization techniques have been applied to areas such as identifying drug interventions from MEDLINE (Fiszman, Demner-Fushman, Kilicoglu, & Rindflesch, 2009).

6.2.5 Visual Design

Interactive functions in CiteSpace correspond to three levels of units of analysis. At the network level, functions operate on networks, including global visualizations of networks: a node-and-link cluster view and a timeline view. At the cluster level, functions operate on individual clusters such as showing all the citers to a cluster or hiding a cluster. At the basic entity level, functions are restricted to individual entities, for example, showing the citation history of a cited reference.

Fig. 6.12 shows a screenshot of the timeline visualization, in which clusters are displayed horizontally alone timelines. In timeline visualizations, the legend above the display area marks every 5 years. The label of each cluster is shown at the end of the cluster's timeline. Cited references or authors are depicted as circles filled with citation rings. The color of each ring corresponds to the time slice in which citations were made. The thickness of a ring is proportional to the amount of citations received in that time slice. Thus, a large-sized circle denotes a highly cited unit, i.e. reference or author. In timeline visualizations of cited authors, a cited author is positioned based on the earliest year in which he/she was cited in the dataset. A possible extension of this design would differentiate citations to the same author in different years.

Two additional colors, red and purple, are used to highlight special attributes of a node. A red ring indicates that a citation burst is detected in the corresponding time slice. A purple ring is added to a node if its betweenness


CiteSpace: Display Merged - (c) 2003-2009 Chaomei Chen - Project Home: C:\Documents and Settings\cchen\Data\Information Science 1996-2008\project
File
Export
Layout
Display
Filtering
Clustering
Help
Spotlight
Citation Burst
Link Walkthrough
# clusters 52
Node Details
MeSH Headings - Major Topics
MeSH Headings - Minor Topics

CiteSpace, v. 2.2.R2
June 18, 2009 2:23:11 PM EDT
C:\Documents and Settings\cchen\Data\Information Science 1996-2008\unique_9478
Timespan: 1996-2008 (Slice Length=1)
Selection Criteria: Top 100 per slice TC=(8, 295)
Network: N=672, E=6071
Modularity Q=0.6798
Mean Silhouette=0.6878

#0 the-digital age (31.65, 1.0E-4)
#1 technological relevance (41.36, 1.0E-4)
#2 nano-titled journal (17.12, 1.0E-4)
#3 academic web site (94.13, 1.0E-4)
#4 everyday life information (11.27, 0.0010)
#5 unique web citation (35.52, 1.0E-4)
#6 south african scholar (28.82, 1.0E-4)
#7 international cooperation (23.63, 1.0E-4)
#8 international cooperation (9.93, 0.0050)
#9 international cooperation (32.02, 1.0E-4)
#10 scientific collaboration (20.56, 1.0E-4)
#11 correspondence factor analysis (15.37, 1.0E-4)
#12 bibliometric output (9.84, 0.0050)
#13 enzyme engineering research (36.93, 1.0E-4)
#14 document type (9.1, 0.0050)
#15 integrating scientometric indicator (25.72, 1.0E-4)
#16 webometric analysis (12.34, 0.0010)
#17 welfare topic (13.14, 0.0010)
#18 knowledge management (19.95, 1.0E-4)
#19 scientific paper (10.67, 0.0050)
#20 different subject field (16.62, 1.0E-4)
#21 stone age baby (23.99, 1.0E-4)
#22 digital library (31.97, 1.0E-4)
#23 stochastic modelling (21.48, 1.0E-4)
#24 generalised informetric analysis (18.37, 1.0E-4)
#25 lotkaian informetric system (14.76, 0.0010)
#26 information organization (10.95, 0.0010)
#27 economic sector (13.14, 0.0010)
#28 scientific journal (25.7, 1.0E-4)
#29 international scientific publication (10.43, 0.0050)
#30 search term (22.18, 1.0E-4)
#31 art history (57.28, 1.0E-4)
#32 physical behavior (53.36, 1.0E-4)
#33 childrens web portal (11.87, 0.0010)
#34 understanding subject heading (12.12, 0.0010)
#35 social capital (15.32, 1.0E-4)
#36 everyday world (12.12, 0.0010)
#37 information behavior (31.37, 1.0E-4)
#38 relevance criteria (46.97, 1.0E-4)
#39 using information retrieval (10.21, 0.0050)
#40 diagnostic problem solving model (14.55, 0.0010)
#41 text grammar (28.09, 1.0E-4)
#42 probabilistic model (91.12, 1.0E-4)
#43 query expansion (41.54, 1.0E-4)
#44 information science (13.24, 0.0010)
#45 cross-language information retrieval (60.26, 1.0E-4)
#46 empirical Model (10.8, 0.0050)
#47 empirical research approache (18.46, 1.0E-4)
#48 utopian place (18.72, 1.0E-4)
#49 h index (116.91, 1.0E-4)
#50 electronic journal (29, 1.0E-4)
#51 high street consumer health information shop (13.65, 0.0010)

Freq	Burst	Centrality	Σ	Keyword	Author	Year	Title	Source	Vol	Page	HalfLife	Cluster
78	6.29	0.04	0.16		LAWRENCE S	1999	Title not found	NATURE	V400	P107	1	3
65	3.16	0.05	0.25		INOWERSEN P	1998	Title not found	J DOC	V54	P236	1	3
63	7.00	0.05	0.17		ALMIND TC	1997	Title not found	J DOC	V53	P404	2	3
57	6.48	0.02	0.11		KLEINBERG JM	1999	Title not found	J ACM	V46	P604	3	3

Fig. 6.12 A timeline visualization of information science. Source: (Chen et al., 2010).

centrality is greater than 0.1; the thickness of the ring is proportional to its centrality value.

A line connecting two items in the visualization represents a co-citation link. The thickness of a line is proportional to the strength of co-citation. The color of a line represents the time slice in which the co-citation was made for the first time. A useful byproduct of spectral clustering is that tightly coupled clusters tend to be placed next to each other and visually form a supercluster.

6.3 A Domain Analysis of Information Science

The knowledge domain in question is information science. A single-slice comparative ACA (2001 – 2005) is described first, followed by progressive ACA and DCA (1996 – 2008). The comparative ACA is compared with the results of the study of Zhao and Strotmann (2008). The two progressive studies analyzed a dataset of 13 years of publications between 1996 and 2008. Within each study, we summarize the prominent co-citation clusters in terms of their leading members, automatically generated labels based on information extracted from citing articles, and sentence summarization based on sentences in citing articles' abstracts.

6.3.1 A Comparative ACA (2001 – 2005)

Zhao and Strotmann (2008) identified 11 specialties of information science based on 120 most cited authors in 2001 – 2005 and manually labeled these specialties by examining each specialty's members. The purpose of the following comparative ACA was to compare with the results the ACA study of Zhao and Strotmann (2008) but using different analytic methods.

We chose the top 120 most cited authors in the same time period using a single 5-year time slice in CiteSpace. Twelve author co-citation clusters were identified with a modularity of 0.5691, suggesting that inter-cluster connections are considerable but not overwhelming. The mean silhouette value of 0.7219 indicates a satisfactory partition of the network. The labels of these clusters were chosen from titles of their citers by *tf*idf* (see Fig. 6.13). In contrast, Zhao and Strotmann (2008) derived their labels from cited authors.

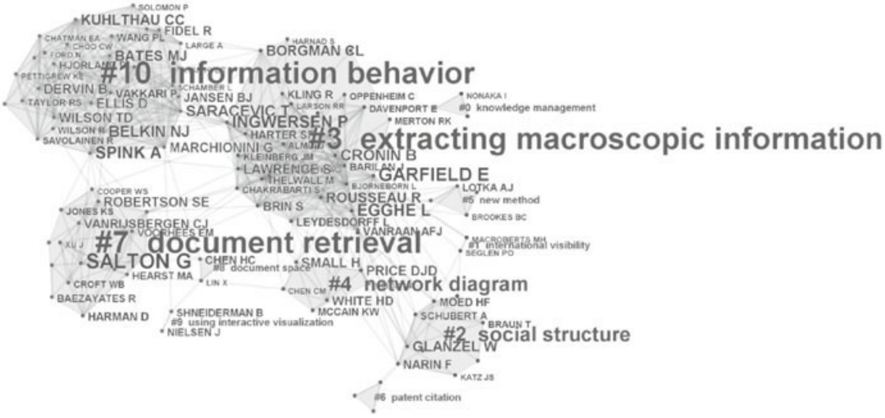


Fig. 6.13 A 120-author ACA network on a single time slice of 5 years (2001 – 2005). Clusters are labeled by citers' title terms using tf*idf weighting. An undefined cluster (#11) is omitted. Source: (Chen et al., 2010).

Determining the number of specialties is a key issue in a co-citation analysis. In factor analysis, it is a common practice to identify specialties in terms eigenvectors with eigenvalues of 1 or greater. Zhao and Strotmann (2008) also took into account other information such as the Scree plot, the total variance explained, communalities and correlation residuals.

We compared 12 co-citation clusters (C_i) and 11 factors (F_j) in Zhao and Strotmann's ACA in terms of their overlapping members. Since each author can only appear in one cluster but may appear in multiple factors, one matching factor was selected only if the author has the greatest factor loading in absolute values; if no such factor was found, the author had no match. The overall overlapping rate is 82%, computed as follows:

$$\frac{\left|\left(\bigcup_{i=1}^{12} C_i\right) \bigcap \left(\bigcup_{j=1}^{11} F_j\right)\right|}{\left|\bigcup_{i=1}^{12} C_i\right|} = \frac{98}{120} = 0.82$$

Each cluster was projected as a distribution of its members over the 11 factors and the no-match category. Cluster C_i's projection on factor F_j is computed as: $\frac{|C_i \bigcap F_j|}{|C_i|}$. For instance, cluster C_3's projection on $F_{\text{webometrics}}$ is $\frac{|C_3 \bigcap F_{\text{webometrics}}|}{|C_3|} = \frac{19}{29} = 0.6552$.

Fig. 6.14 depicts the overall matching patterns between clusters (diamonds) and factors (circles) in a similarity graph. The thickness (and darkness) of a line indicates the strength of the match. There are three types of patterns.

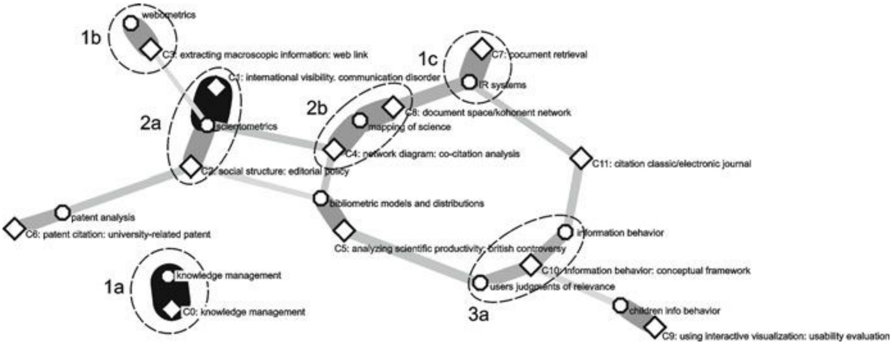

1b
webometrics
C3: extracting macroscopic information: web link
1c
C7: document retrieval
C1: international visibility. communication disorder
IR systems
2b
C8: document space/kohonent network
2a
scientometrics
mapping of science
C4: network diagram: co-citation analysis
C11: citation classic/electronic journal
C2: social structure: editorial policy
bibliometric models and distributions
patent analysis
C6: patent citation: university-related patent
information behavior
C5: analyzing scientific productivity: british controversy
C10: information behavior: conceptual framework
1a
knowledge management
users judgments of relevance
C0: knowledge management
3a
children info behavior
C9: using interactive visualization: usability evaluation

Fig. 6.14 An associative network of clusters (diamonds) and factors (circles) with 10% or more overlaps (thickness of line). Cluster labels are shown in two parts: terms chosen by tf*idf and by log-likelihood ratio. Source: (Chen et al., 2010).

Type 1a-1c: a cluster primarily corresponds to a single factor, denoted as $C_i \Leftrightarrow F_j$.

(1a) $C_0 \Leftrightarrow F_{\text{knowledge management}}$

(1b) $C_3 \Leftrightarrow F_{\text{webometrics}}$

(1c) $C_7 \Leftrightarrow F_{\text{IR systems}}$

Type 2a-2b: two or more clusters are subsets of the same factor, i.e. for K clusters, $\bigcup_{k=1}^{K} C_{i_k} \subseteq F_j$.

(2a) $C_1 \bigcup C_2 \bigcup C_4 \subseteq F_{\text{scientometrics}}$

(2b) $C_4 \bigcup C_8 \subseteq F_{\text{mapping of sciencs}}$

Type 3a: one cluster is split into L factors, i.e $\bigcup_{l=1}^{L} F_{jt} \subseteq C_i$.

(3a) $F_{information\ behavior} \cup F_{users\ judgements\ of relevance}$
$\cup F_{childrens\ information\ behavior} \subseteq C_{10}$

Factor labels given by Zhao and Strotmann tend to be conceptually higher-level terms than automatically generated cluster labels. For example, patent analysis is a broader term of patent citation; and IR systems is a broader term of document retrieval. Structurally, spectral clusters tend to be more specific groupings than factors. For instance, as shown in 2b, the mapping of science factor contains clusters such as C_4*: network diagram* (by tf*idf): co-citation analysis (by LLR), and C_8: document space/Kohonen network.

In summary, (1) spectral clustering and factor analysis identified about the same number of specialties, but they appeared to reveal different aspects of co-citation structures, and (2) cluster labels chosen from citers of a cluster tend to be more specific terms than those chosen by human experts. These findings suggest that the multiple perspective method has the potential to provide additional insights in complementary to existing methods and provide an intermediate level of support for interpreting the nature of specialties.

6.3.2 A Progressive ACA (1996–2008)

This progressive ACA is a 13-year multiple-slice analysis of 5,963 records of articles and reviews. A progressive co-citation analysis takes multiple co-citation networks from consecutive time intervals as input and produces a merged network to represent the evolution of the underlying domain (Chen, 2004). The inclusion of review-type records was to cover ARIST publications, which are classified as reviews. By including top-150 most cited authors from every year between 1996 and 2008, we obtained a merged network of 633 cited authors with 7,162 author co-citation links and 40 co-citation clusters. This 633-author network has a lower modularity (0.2278) than the smaller 120-author network in the comparative ACA (0.5691). Furthermore, the mean silhouette value (0.6929) of the larger network is also lower than that of the 120-author network (0.7219). The larger network has a much higher inter-cluster connectivity.

A timeline visualization of the 40 ACA clusters is shown in Fig. 6.15 with automatically generated cluster labels. The display shows labels of highly cited authors in major clusters only. These clusters appear to be more interpretable than clusters in the comparative ACA. Visually, one may identify a few superclusters at the granularity of Zhao and Strotmann's factors. For example, clusters C_2, C_4, C_7, and C_8 form a supercluster that corresponds to the *scientometrics* factor.

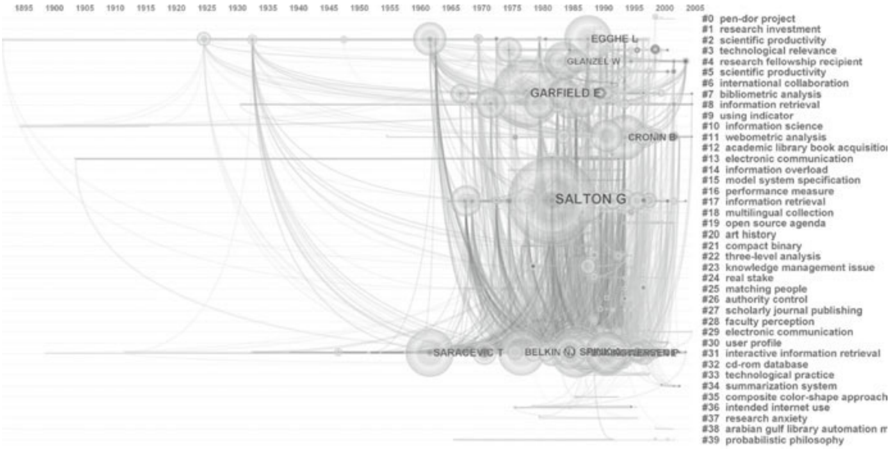


Fig. 6.15 40 ACA clusters (1996 – 2008) (Nodes=633, Edges=7,162, top N=150, time slice length=1, modularity=0.2278, mean silhouette=0.6929). Source: (Chen et al., 2010).

Table 6.1 shows automatically chosen cluster labels of the 6 largest ACA clusters along with their size and silhouette value. Top-ranked title terms by LLR were selected as cluster labels. The largest cluster *interactive information retrieval* (#31) has 199 members. Its negative silhouette value of −0.090 suggests a heterogeneous citer set. The second largest cluster (#17), with 95 members, is labeled as *information retrieval.* Other candidate labels for the cluster include *probabilistic model* and *query expansion*, confirming that this cluster deals with classic information retrieval issues. The third largest cluster (#7) is *bibliometric analysis.*

Most cited authors include Spink_A and Saracevic_T in *interactive information retrieval* (#31), Salton_G, Robertson_SE, and van Rijsbergen_CJ in *information retrieval* (#17), Garfield_E, Moed_HF, and Merton_RK in *bibliometric analysis* (#7), Egghe_L, Price_DJD, and Lotka_AJ in *statistical analysis* (#2). The *webometric analysis* cluster (#11) includes Cronin_B, Rousseau_R, and Lawrence_S. The *journal co-citation analysis* (#8) includes Small_H, Leydesdorff_L, and White_HD.

Note that one may reach different insights into the nature of a co-citation cluster if different sources of information are used. The cited members of a cluster define its intellectual base, whereas citers to the cluster form a research front. The major advantage of our approach is that it enables analysts to consider multiple aspects of the citation relationship from multiple perspectives.

Table 6.1 6 largest ACA clusters of a 633-author network (1996 – 2008).

$C\#$	n	Silhouette	Title terms by tf*idf	Title terms by LLR ($p = 0.0001$)
31	199	−0.090	(80.30) interactive information retrieval (62.46) information retrieval (55.45) information science	(79.31) **interactive information retrieval** (53.64) user information problem (53.64) various aspect
17	95	0.143	(71.02) information retrieval (41.44) probabilistic model (37.94) query expansion	(68.19) information retrieval (49.81) probabilistic model (38.49) magazine article
7	37	0.272	(22.87) bibliometric analysis (17.07) social science (13.82) publication productivity	(31.51) **bibliometric analysis** (22.79) social science (22.67) career path
2	33	0.343	(20.97) scientific productivity (14.98) statistical analysis (14.76) analyzing scientific productivity	(24.24) theoretical population genetic (24.16) **statistical analysis** (23.54) new method
11	32	0.682	(20.72) webometric analysis (18.44) informetric purpose (18.44) data collection method	(32.65) **webometric analysis** (28.16) data collection method (28.16) informetric purpose
8	30	0.330	(19.68) information retrieval (18.42) **citation analysis** (16.12) information retrieval area	(34.42) **journal co-citation analysis** (34.42) intellectual space (34.42) information retrieval area

6.3.3 A Progressive DCA (1996 – 2008)

In the progressive DCA, co-citation networks were first constructed with the top-100 most cited documents in each of the 13 one-year time slices between 1996 and 2008. Then, these networks were merged into a network of 655 co-cited references. The merged network was subsequently decomposed into 50 clusters. Table 6.2 summarizes these clusters. We first provide an overview of these clusters and discuss the five largest clusters in detail.

Table 6.2 The five largest clusters sorted by size.

$C\#$	n	%	Silhouette	Title terms (tf*idf)	Title terms (LLR) (*p=0.0001)
18	150	22.90	−0.024	(80.82) **interactive information retrieval** (72.92) information retrieval (51.5) user information problem	**interactive information retrieval** (294.13*) user information problem (167.06*) various aspect (167.06*)
43	69	10.53	0.522	(131.97) **academic web** (103.62) web site (54.77) exploratory hyperlink	academic web (174.79*) exploratory hyperlink (152.94*) linguistic consideration (152.94*)
13	46	7.02	0.153	(42.16) information retrieval (23.47) probabilistic model (22.53) query expansion	**information retrieval** (104.03*) probabilistic model (81.54*) using heterogeneous thesauri (67.95*)
35	44	6.72	0.245	(14.07) **citation behavior** (14.07) citing literature (11.74) citation theory	**citation behavior** (56.66*) citing literature (56.66*) citation theory (43.92*)
2	29	4.43	0.834	(83.69) **h-index** (53.18) successive h-indices (43.03) generalized hirsch h-index	**h-index** (212.76*) generalized hirsch h-index (156.63*) disclosing latent fact (156.63*)

The 50 clusters vary considerably in size. The largest cluster #18 contains 150 members, which is 22.90% of the entire set of 655 references. The five largest clusters altogether reach 51.60%. In contrast, there are six clusters contain only two members.

The network's overall mean silhouette value is 0.7372, which is the highest among the three co-citation networks we analyzed in this study. In general, the silhouette value of a cluster is negatively correlated with its size (−0.654). For example, the largest cluster, #18, has the lowest silhouette value of −0.024, indicating its diverse and heterogeneous structure. In contrast, the second largest cluster, #43, has a more homogenous structure with a reasonably high silhouette value of 0.522. The fifth largest cluster, #2, has a very high silhouette value of 0.834. The following discussion will focus on the five largest clusters and their interrelationships.

The five largest document co-citation clusters are interactive information retrieval (#18), academic web (#43), information retrieval (#46), citation behavior (#44), and h-index (#2). We analyzed two aspects of each specialty: (1) prominent members of a cluster as the intellectual basis and (2) themes identified in the citers of the cluster as research fronts.

Table 6.3 summarizes two clusters (#43 and #2, both have silhouette values greater than 0.50) in terms of top-cited members and their structural, temporal, and saliency metrics such as citation count (φ), betweenness centrality (σ), citation burstness (τ), and sigma — a novelty indicator ($\sum$) (Chen et al., 2009a). The stars in the academic web cluster (#43) are Lawrence_1999 and Kleinber_1999. Both papers were published outside the domain defined by the 12 source journals; instead, they appeared in *Nature* and JACM. This is an example of how one discipline (information science) was influenced by another (computer science). The core of the fifth largest cluster, the h-index cluster (#2), is Hirsch_2005, which originally introduced the concept of h-index. The strongest citation burst of 15.75 was detected in the citation history of Hirsch_2005. As our analysis will demonstrate, the h-index cluster is one of the most active areas of research in recent years.

Table 6.3 Most frequently cited references in two of document co-citation clusters.

Cluster #	φ	τ	σ	$\sum$	Cited references
43	76	8.83	0.06	0.17	LAWRENCE S (1999) Accessibility and distribution of information on the Web, Nature, 400, 107.
	64	9.00	0.06	0.16	ALMIND TC (1997) Informetric analyses on the world wide web: methodological approaches to 'Webometrics', J DOC, 53, 404.
	63	3.44	0.04	0.22	INGWERSEN P (1998) The calculation of Web impact factors. J DOC, 54, 236
	53	6.58	0.02	0.12	Kleinberg, J. M. (1999) Authoritaive sources in a hyperlinked environment. JACM, 46, 604-632.
	50	7.12	0.03	0.13	Rob Kling and Geoffrey W. McKim (2000) Not just a matter of time: Field differences and the shaping of electronic media in supporting scientific communication. JASIS, 51(14), 1306-1320.
2	42	15.75	0	0.02	HIRSCH JE (2005) An index to quantify an individual's scientific research output, P NATL ACAD SCI USA, 102, 16569
	24	8.98	0	0.01	Bornmann, L. & Daniel, H.-D. (2005) Does the *h*-index for ranking of scientists really work? Scientometrics, 65(3), 391-392.
	22	7.54	0	0.02	Ball, P. (2005) Index aims for fair ranking of scientists, NATURE, 436(7053), 900.
	19	7.11	0	0.01	Branu, Tibor (2005) A Hirsch-type index for journals, The Scientists, 19(22), 8.
	18	6.73	0	0.02	Egghe, L. (2005). Power laws in the information production process: Lotkaian informetrics. Elsevier: Oxford, UK.

The average age of core papers in a cluster is an estimation of the time the cluster was formed. According to the average age of top-5 core papers, the 37-year old citation cluster is the oldest — formed around 1973, its average year of publication, whereas the h-index cluster is the youngest — 5 years old, formed in 2005. In between, the Information Retrieval cluster (#13) is 31 (formed in 1979); the interactive Information Retrieval cluster (#18) is 18 formed in 1992; and the academic web cluster (#43) is 11 (formed in 1999).

Research fronts of a document co-citation cluster were characterized by terms extracted from the citers of the cluster. Nine methods of ranking extracted terms were implemented in CiteSpace by choosing terms from three sources — titles, abstracts, and index terms of the citers of each cluster — and three ranking algorithms, namely, tf*idf weighting (Salton et al., 1975), log-likelihood ratio tests (LLR) (Dunning, 1993; Witten & Frank, 1999), and mutual information (MI)(Witten & Frank, 1999). Top-ranked terms became candidate cluster labels.

The reliability of these term ranking methods was measured by a consensus score $r = 0.1 * (n + 1)$, where n is the number of other methods that also top rank the same term. It turned out that the best three ranking methods were: (1) title terms ranked by LLR, (2) index terms ranked by LLR, and (3) title terms ranked by tf*idf. tf*idf and LLR produced identical labels for 36 clusters out of 50 (72%).

The largest cluster (#18) has 150 members and it has the lowest silhouette value. It turned out that the cluster was cited by 185 citing articles in the dataset. A total of 869 terms were extracted from the titles of these citing articles. In order to verify the heterogeneity of this set of citers, the term similarity network was decomposed using singular value decomposition (SVD). As a result, the term space was indeed multi-dimensional in nature because the largest connected component of the term similarity network contains only 353 terms, which is 40.62% of the 869 terms. In contrast, the h-index cluster (#2) was much more homogeneous; the cluster was the citation footprint of 39 citing articles.

The second largest cluster was labeled as academic web by LLR, but the top-ranked index term was webometrics, which was also the name of a specialty identified by Zhao and Strotmann. The index term webometrics is broader and more generic than the term academic web. This observation suggests that a manual labeling process is probably very similar to the indexing process after all.

The identification of the h-index cluster is unique because there was no such cluster in the 1996 – 2008 ACA. This is a good example why one should consider both ACA and DCA so that distinct DCA clusters such as the h-index one can be detected.

The time span τ between a research front and its intellectual base can be estimated as the difference between their average years of publications:

$$\tau(C_i) = \frac{\sum_{d \in citers(C_i)} year(d)}{|citers(C_i)|} - \frac{\sum_{c \in C_i} year(c)}{|C_i|} + 1$$

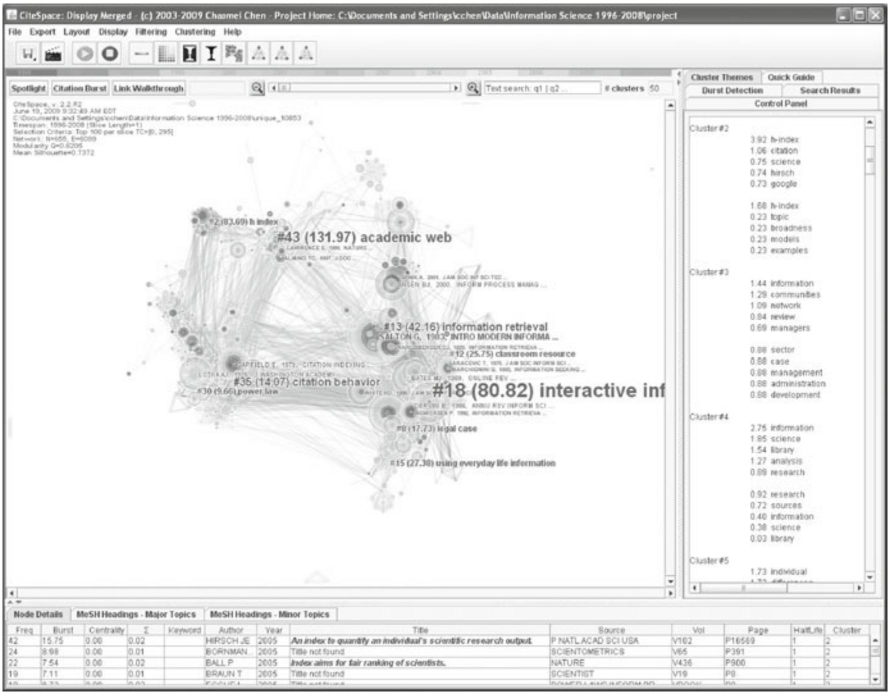
For example, citation behavior (#35) has the longest time span, $\tau(C_{35})$ = 2000-1973=28 years. IR has the second longest time span $\tau(C_{13})$ = 2000-1979=22. The time span for interactive IR is $\tau(C_{18})$ = 2000-1992=9 years; for academic web (#43), $\tau(C_{43})$ = 2003-1999=5 years; and for h-index (#2), $\tau(C_2)$ = 2007-2005=3 years.

Table 6.4 lists the most representative citing articles in each cluster. For example, Thelwall has a prominent role in the research front of the academic web cluster (#43). He authored 3 of the top 5 citing articles of the cluster, including Thelwall_2003 which cited 14 references of the cluster.

Table 6.4 Titles of the two most frequent citers to each of the 5 largest DCA clusters. Terms chosen by LLR are underlined.

#	Cluster label	Titles of key citers
18	Interactive information retrieval	(16) Robins D (2000) shifts of focus on various aspects of user information problems during interactive information retrieval (15) Beaulieu M (2000) interaction in information searching and retrieval
43	Academic web	(14) Thelwall M (2003) disciplinary and linguistic considerations for academic web linking: an exploratory hyperlink mediated study with mainland china and taiwan (12) Wilkinson D (2003) motivations for academic web site interlinking: evidence for the web as a novel source of information on informal scholarly communication
13	Information retrieval	(8) Ding Y (2000) bibliometric information retrieval system (birs): a web search interface utilizing bibliometric research results (6) Dominich S (2000) a unified mathematical definition of classical information retrieval (6) Sparck-Jones K (2000) a probabilistic model of information retrieval: development and comparative experiments part 2
35	Citation behavior	(5) Case DO (2000) how can we investigate citation behavior? a study of reasons for citing literature in communication (5) Ding Y (2000) bibliometric information retrieval system (birs): a web search interface utilizing bibliometric research results
2	H-index	(14) Bornmann L (2007) what do we know about the h-index? (11) Sidiropoulos A (2007) generalized hirsch h-index for disclosing latent facts in citation networks

The DCA network shown in Fig. 6.16 was generated by CiteSpace. The 655 references and 6,099 co-citation links were divided into 50 clusters with a modularity of 0.6205, which represents a considerable amount of inter-cluster links. Major clusters are labeled in the visualization in red color with the font size proportional to the size of clusters. The colors of co-citation links reveal that the earliest inter-cluster connection is between interactive IR and IR,



CiteSpace: Display Merged - (c) 2003-2009 Chaomei Chen - Project Home: C:\Documents and Settings\cchen\Data\Information Science 1996-2008\project

CiteSpace, v. 2.2.R2
June 19, 2009 9:32:49 AM EDT
C:\Documents and Settings\cchen\Data\Information Science 1996-2008\unique_10853
Timespan: 1996-2008 (Slice Length=1)
Selection Criteria: Top 100 per slice TC=[0, 295]
Network: N=655, E=6099
Modularity Q=0.8205
Mean Silhouette=0.7372

Cluster #2
3.92 h-index
1.06 citation
0.75 science
0.74 hirsch
0.73 google

1.68 h-index
0.23 topic
0.23 broadness
0.23 models
0.23 examples

Cluster #3
1.44 information
1.29 communities
1.09 network
0.84 review
0.69 managers

0.88 sector
0.88 case
0.88 management
0.88 administration
0.88 development

Cluster #4
2.76 information
1.85 science
1.54 library
1.27 analysis
0.89 research

0.92 research
0.72 sources
0.40 information
0.38 science
0.03 library

Cluster #5
1.73 individual

Freq	Burst	Centrality	Σ	Keyword	Author	Year	Title	Source	Vol	Page	HalfLife	Cluster
42	15.75	0.00	0.02		HIRSCH JE	2005	***An index to quantify an individual's scientific research output***	P NATL ACAD SCI USA	V102	P16569	1	2
24	8.98	0.00	0.01		BORNMAN...	2005	Title not found	SCIENTOMETRICS	V65	P391	1	2
22	7.54	0.00	0.02		BALL P	2005	***Index aims for fair ranking of scientists.***	NATURE	V436	P900	1	2
19	7.11	0.00	0.01		BRAUN T	2005	Title not found	SCIENTIST	V19	P8	1	2

File
Export
Layout
Display
Filtering
Clustering
Help
Spotlight
Citation Burst
Link Walkthrough
Text search: q1 | q2...
# clusters 50
Cluster Themes
Quick Guide
Burst Detection
Search Results
Control Panel
Node Details
MeSH Headings - Major Topics
MeSH Headings - Minor Topics
#2 (83.69) h index
#43 (131.97) academic web
LAWRENCE S, 1999, NATURE ...
#13 (42.16) information retrieval
SALTON G, 1983, INTRO MODERN INFORMA ...
#12 (25.75) classroom resource
GARFIELD E, 1979, CITATION INDEXING ...
#35 (14.07) citation behavior
#30 (9.66) power law
#18 (80.82) interactive inf
#8 (17.73) legal case
#15 (27.38) using everyday life information

Fig. 6.16 An overview of the co-citation networks. Cited references with highest sigma values are labeled. Source: (Chen et al., 2010).

followed by connections between interactive IR and citation behavior, then by the more recent connections between academic web, citation behavior, and IR, and finally, the most recent connections between h-index, citation behavior, academic web.

Fig. 6.17 shows a timeline visualization of the 50 clusters and their interrelationships. Each cluster is plotted horizontally. Each timeline runs from left to right with its label displayed to the right. The design of the timeline visualization is inspired by the work of Morris, Yen, Wu, and Asnake (2003) and further enhanced by automatic clustering labeling with multiple algorithms. Analysts can visually identify a variety of characteristics of a cluster, such as the length of its history, its citation classics, citation bursts, and how it is connected to other clusters. For example, in the citation behavior cluster, Garfield's 1979 book on citation index stands out with the highest betweenness centrality of 0.09. Many large citation circles would identify a high-impact specialty, whereas many red rings of citation bursts would highlight emerging specialties.

It is clear in the timeline visualization that the h-index cluster is new and growing fast. The new cluster includes not only a highly cited article, but also with a strong surge of citations. Hirsch_2005, the original article that introduces the h-index, has the strongest citation burst (see Fig. 6.18).

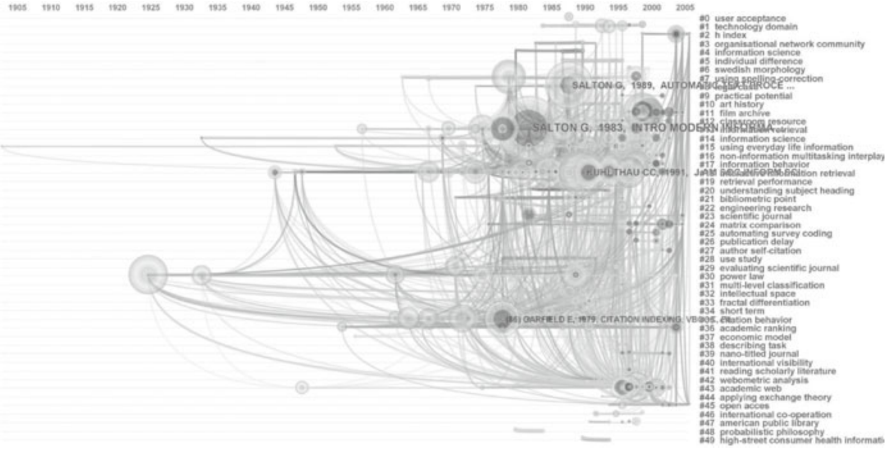


Fig. 6.17 A timeline visualization of the 50 DCA clusters (655 nodes, 6,099 links, modularity=0.6205, mean Silhouette=0.7372). Cluster labels are automatically generated from title terms of citing articles of specific clusters. Source: (Chen et al., 2010).

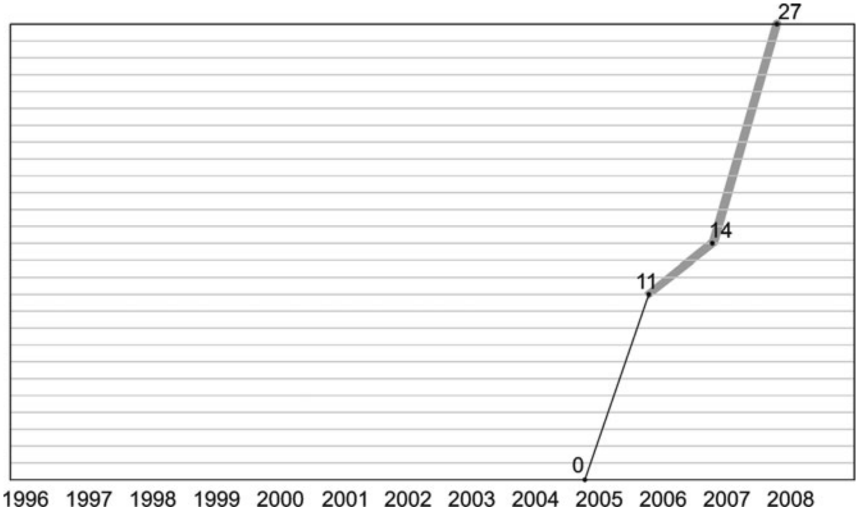


Fig. 6.18 The burst of citations to Hirsch_2005. Source: (Chen et al., 2010).

If we start from the top of the timeline visualization and move down line by line, we can see many representative references in these clusters. For example, Film Archive (#11) is a relatively new cluster, containing Jansen_2000 on searching for multimedia on the web as a major reference. Similarly, the most cited reference in information retrieval (#13) is Salton's book. Further down the timeline list is the citation behavior cluster (#35), which features

Garfield_1979 prominently. Many co-citation links join citation behavior and academic web. Some long-range co-citation links connect the h-index cluster and other clusters such as the academic web cluster and the power law cluster.

6.4 Summary

Progressive knowledge domain visualization represents an extension of the traditional citation-based analysis of scientific literature along the temporal dimension. A time series of snapshots of a domain are synthesized so that critical paths of the evolution of the domain can be found. The multiple-perspective approach broadens the traditional approaches from a shadow-focused perspective to multiple perspectives so that analysts can obtain multiple views of the same domain.

The 5-year comparative ACA reveals that human experts tend to choose broader-term labels, equivalent to the level of index terms, whereas algorithmically chosen terms from titles or abstracts tend to be more specific and limited to terms actually used by the authors. The 13-year progressive ACA (1996 – 2008) of 633 cited authors in 40 clusters resulted in a clearer global structure than the 5-year single-slice ACA. We have shown that cluster membership and citer-focused labeling provide complementary information to form a comprehensive image of the field. The progressive DCA of 655 cited references detects a distinct and fast growing cluster — the h-index cluster, which is absent from the progressive ACA. The h-index cluster emerged in 2005.

The comparison with the study of Zhao and Strotmann was very valuable. It offered us an opportunity to compare the analysis conducted by human experts to the interpretation cues provided by our automatic labeling and summarization methods. We did not include an investigation of direct citation networks in our study. Some recent work in this area can be found in (Garfield, 2004; Morris & Van der Veer Martens, 2008). These alternative studies should be considered and compared thoroughly in the future work.

The quality of cluster labeling in general depends on the variety, breadth, and depth of the set of candidate terms. We have extracted candidate cluster labels from citing articles' titles and abstracts. It is also desirable to compare these labels with candidate terms from cited references in the DCA or cited authors' publications in ACA. However, such data is not readily available due to the fact that the CR field in the ISI data format does not include the title of a cited reference. Furthermore, a cited reference may not even be a source record in the entire collection of the Web of Science. In contrast, human analysts may choose the most appropriate label terms from a much wider range of sources beyond the terms found in the immediate dataset, although this could be a double-edged sword. On the one hand, human analysts are free from the limitations of a specific data source. On the other hand, they may need to deal with a potentially much larger search space, which can be a daunting task, especially for those who do not have an encyclopedic knowledge of the subject domain. Utilizing external information sources such as the Wikipedia and the World Wide Web is a promising direction to resolve the problems due to the limited term space problem. An interesting approach was reported recently in (Carmel, Roitman, & Zwerdling, 2009).

Although some cluster labels make good sense, some labels are still puzzling and some members of clusters may not be as intuitive as others. Some of the labels appear to be strongly biased by particular citing articles, especially when the size of a cluster is relatively small. Algorithmically generated cluster labels are limited to deal with clusters that have multiple aspects formed by a diverse range of citing papers. Clusters with low mean silhouette values tend to be subject to such limitations more than high silhouette clusters. On a positive note, metrics such as modularity and silhouette provide useful indicators of uncertainty that analysts should take into account when interpreting the nature of clusters. We have been looking for a labeling algorithm that is consistently better than others. Since we do not have datasets with gold standards, this cannot be validated systematically except by making comparisons across the 9 sets of candidate labels.

A few more fundamental questions need to be thoroughly addressed. If labels selected from citers differ from those from citees, how do we reconcile the difference? How do we make sense of the citer-citee duality? One of the fundamental assumptions for co-citation analysis is that co-citation clusters do represent something substantial as real as part of the reality although they might be otherwise invisible. Given that some co-citation clusters appear to be biased by the citation behavior of particular publications, it may become necessary to re-examine the assumption, especially whether co-citation clusters represent something that is truly integral to the scientific community as a whole.

The multiple-perspective approach has the following advantages over the traditional one:

1) It can be consistently used for both DCA and ACA.
2) It uses more flexible and efficient spectral clustering to identify co-citation clusters.
3) It characterizes clusters with candidate labels selected by multiple ranking algorithms from the citers of these clusters and reveals the nature of a cluster in terms of how it has been cited.
4) It provides metrics such as modularity and silhouette as quality indicators of clustering to aid interpretation tasks.
5) It provides integrated and interactive visualizations for exploratory analysis.

These features enhance the interpretability and accountability of co-citation analysis. Modularity and silhouette metrics provide useful quality indicators of clustering and network decomposition. This is a valuable addition to the traditional methods such as estimating the strength of a membership based on factor loading. Multiple channels of candidate labels selected from multiple sources have confirmed that clustering labeling is indeed a complex phenomenon that requires multiple perspectives.

The integration of these techniques in a unifying framework will enable analysts, researchers, and students to investigate and understand the dynamic interrelationship between an intellectual base and an inspired research front. Multiple perspective approaches also provide a cross-validation basis for evaluation and comparison.

References

Bonacich, P. (1987). Power and centrality: A family of measures. American Journal of Sociology, 92, 1170-1182.

Brandes, U. (2001). A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 25(2), 163-177.

Bursill, H. (1859). Hand shadows to be thrown upon the wall. Griffith and Farran.

Carmel, D., Roitman, H., & Zwerdling, N. (2009). Enhancing cluster labeling using wikipedia. Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (pp. 139-146).

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. Proc. Natl. Acad. Sci. USA, 101(suppl), 5303-5310.

Chen, C. (2005). The centrality of pivotal points in the evolution of scientific networks. Proceedings of the international conference on intelligent user interfaces (IUI 2005) (pp. 98-105). ACM Press.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57(3), 359-377.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009a). Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics, 3(3), 191-209.

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. Journal of the American Society for Information Science and Technology, 61(7), 1386-1409.

Chen, C., & Kuljis, J. (2003). The rising landscape: A visual exploration of superstring revolutions in physics. Journal of the American Society for Information Science and Technology, 54(5), 435-446.

Chen, C., Zhang, J., & Vogeley, M.S. (2009b). Mapping the global impact of Sloan digital sky survey. IEEE Intelligent Systems, 24(4), 74-77.

Cronin, B. (1981). Agreement and divergence on referencing practice. Journal of Information Science, 3(1), 27-33.

Deerwester, S., Dumais, S.T., landauer, T.K., Furnas, G.W., & Harshman, R.A. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), 61-74.

Fiszman, M., Demner-Fushman, D., Kilicoglu, H., & Rindflesch, T.C. (2009). Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. Journal of Biomedical Informatics, 42, 801-813.

Freeman, L.C. (1977). A set of measuring centrality based on betweenness. Sociometry, 40, 35-41.

French, B.M., & Koeberl, C. (2010). The convincing identification of terrestrial meteorite impact structures: What works, what doesn't, and why. Earth-Science Reviews, 98, 123-170.

Garfield, E. (1979). Citation indexing: Its theory and applications in science, technology, and humanities. New York: John Wiley.

Garfield, E. (2004). Historiographic mapping of knowledge domains literature. Journal of Information Science, 30(2), 119-145.

Jaccard, P. (1901). Éude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin del la Société Vaudoise des Sciences Naturelles, 37, 547-579.

Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graphs. Information Processing Letters, 31(1), 7-15.

Kiss, C., & Bichler, M. (2008). Identification of influencers: Measuring influence in customer networks. Decision Support Systems, 46(1), 233-253.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. Proceedings of Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 91-101). ACM Press.

Kumar, R., Novak, J., Raghavan, P., & Tomkins, A. (2003). On the bursty evolution of blogspace. Proceedings of WWW2003 (pp. 568-576). ACM Press.

Luxburg, U.v. (2006). A tutorial on spectral clustering. http://www.kyb.mpg.de/publications/attachments/Luxburg06_TR_%5B0%5D.pdf. Accessed 1 Oct 2002.

Luxburg, U.v., Bousquet, O., & Belkin, M. (2009). Limits of spectral clustering, from < http://kyb.mpg.de/publications/pdfs/pdf2775.pdf >

Morris, S.A., & Van der Veer Martens, B. (2008). Mapping research specialties. Annual Review of Information Science and Technology, 42, 213-295.

Morris, S.A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. Journal of the American Society for Information Science and Technology, 54(5), 413-422.

Newman, M.E.J. (2006). Modularity and community structure in networks. PNAS, 103(23), 8577-8582.

Ng, A.Y., Jordan, M.I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. Advanced in Neural Information Processing Systems, 14(2), 849-856.

Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Salton, G., Yang, C.S., & Wong, A. (1975). A vector space model for information retrieval. Communications of the ACM, 18(11), 613-620.

Schneider, J.W. (2009). Mapping of cross-reference activity between journals by use of multidimensional unfolding: Implications for mapping studies. In B. Larsen & J. Leta (Eds.), Proceedings of 12th international conference on scientometrics and informetrics (ISSI 2009) (pp. 443-454). BIREME/PAHO/WHO and Federal University of Rio de Janeiro.

Schvaneveldt, R.W. (Ed.). (1990). Pathfinder associative networks: Studies in knowledge organization. Norwood, New Jersey: Ablex Publishing Corporations.

Schwarz, J.H. (1982). Superstring theory. Physics Reports-Review Section of Physics Letters, 89(3), 224-322.

Schwarz, J.H. (1996). The second superstring revolution. http://arxiv.org/PS_cache/hep-th/pdf/9607/9607067.pdf. Accessed 1 Oct 2002.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8), 888-905.

Shibata, N., Kajikawa, Y., Taked, Y., & Matsushima, K. (2008). Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation, 28(11), 758-775.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. Journal of the American Society for Information Science, 24, 265-269.

Small, H. (1986). The synthesis of specialty narratives from co-citation clusters. Journal of the American Society for Information Science, 37(3), 97-110.

Small, H.G. (1977). A co-citation model of a scientific specialty: A longitudinal study of collagen research. Social Studies of Science, 7, 139-166.

Sparck Jones, K. (1999). Automatic summarizing: Factors and directions. In I. Mani & M.T. Maybury (Eds.), Advances in automatic text summarization (pp. 2-12). Cambridge, MA: MIT Press.

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. Computational Linguistics, 28(4), 409-445.

White, H.D. (2007a). Combining bibliometrics, information retrieval, and relevance theory, Part 1: First examples of a synthesis. Journal of the American Society for Information Science and Technology, 58(4), 536-559.

White, H.D. (2007b). Combining bibliometrics, information retrieval, and relevance theory, Part 2: Some implications for information science. Journal of the American Society for Information Science and Technology, 58(4), 583-605.

Witten, I.H., & Frank, E. (1999). Data mining: Practical machine learning tools and techniques with Java implementations. San Francisco, CA: Morgan Kaufmann.

Zhao, D.Z., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996 – 2005: Introducing author bibliographic-coupling analysis. Journal of the American Society for Information Science and Technology, 59(13), 2070-2086.

Chapter 7 Messages in Text

Text is critical in our communication. An overwhelmingly portion of the text we deal with everyday is unstructured: key terms in text are usually not marked as such explicitly and essential claims in text documents are usually not highlighted by colorful inks. Human readers are capable of identifying and discerning implicit messages from text expressed in their natural languages. However, human readers' ability is limited when facing hundreds, thousands, and even more such identification and differentiation tasks. Many real-life situations require a timely understanding of a pile of text or a swift analysis of multiple sources of text. For example, funding agencies need to deal with increasing numbers of proposals; scientists need to keep abreast of many seemingly relevant new publications to their research all the time; and historians need to sift through mountains of archival documents.

7.1 Differentiating Conflicting Opinions

Conflicting opinions are part of the life. At a larger scale, debates can last for years about the causes of mass extinctions hundreds of million years ago. Debates like the one concerning the competitiveness of science and technology in the Gathering Storm can involve a wide range of stakeholders and decision makers. At a smaller scale, reviewers may give contradictory recommendations on whether particular research proposals should be funded. Consumers may find drastically different opinions on whether a new book or a new product is worth purchasing. These types of clashes of opinions are an essential and valuable driving force in situational awareness and decision making. As we have seen from the examples discussed in Chapter 2, contradictions are often an integral part of creativity.

Critical challenges are to identify the basic premises of arguments from each individual perspective, assess the credibility of available evidence and alternative perspectives, understand the context and background of a particular position, and track the development of how various perspectives in a broad context over a long period of time. While detecting trends and dynam-

C. Chen, *Turning Points*

ics of change attracts an increasing interest, fundamental challenges remain at both macroscopic and microscopic levels due to the dynamic and complex nature of our perception and cognition (Thomas & Cook, 2005). While techniques such as topic modeling are powerful in handling a large amount of text, the burden of interpretation and diagnosis essentially remains on the end users, i.e. the analysts. For instance, given a set of polarized customer reviews, the key questions include how many subtopics are being debated, what is the nature of each of the subtopics, and how exactly the conflicting opinions differ. In a more generic form, the question is if the system can identify an emerging trend, how does it differ in precise terms from existing trends?

Table 7.1 illustrates the challenge. The table shows the most prominent topic identified by topic modeling from positive, neutral, and negative reviews of a bestseller book. The words in italic are unique to each subtopic. We can see trees, but not how these trees forming a forest.

Table 7.1 Words in a topic identified by topic modeling from each category of reviews.

Positive	Neutral	Negative
book read reading story *put great* books good brown *dan* time *page ve* interesting don *recommend find found thought*	book brown reading story good *plot code vinci* characters *da fiction* history reading dan *holy grail* author interesting don	book read brown *characters plot* story *writing* time *author* good interesting reading don *people written reader make history character*

At the topic level, each topic is typically represented by an array of words. It is hard for users to make sense precisely which statements or arguments contribute the most to the emerged topic. While a practical representation often seen in reporting the results of topic modeling is to present a passage and highlight words in the passage in corresponding to the generative topics that characterize the host document, there is a lack of a trail of evidence that the user can move back and forth between the topic-level representations and text-level representations.

Ideal representations of evidence for interpreting and making sense of an emerging topic should allow analysts to examine not only patterns generated by statistical and machine-learning algorithms, but also evidence in concrete terms that distinguish patterns of different categories such as positive and negative customer reviews or revolutionary and evolutionary trends.

To meet these criteria, we propose that evidence can be more meaningfully represented in a new kind of a decision tree. Such representations will provide abstractions below the topic-level representations of the overall corpus, but above the annotated text of the original source. Therefore, decision-tree representations will provide an intermediate layer to facilitate the navigation in both directions. In a similar way, association rules of more specific diagnostic values can be incorporated to enrich summaries of competing evidence and help users to understand what exactly the nature of an identified new thematic pattern.

7.1.1 The Da Vinci Code

We use a study of positive and negative reviews to illustrate technical challenges and analytical tasks involved in differentiating conflicting opinions. Reviews of a controversial bestseller book such as *The Da Vinci Code* carry the hallmark of conflicting opinions (Fig. 7.1). Reviews made by readers from a diverse range of perspectives provide a valuable source of insight in terms of how people tend to form their opinions and what influences their opinions. Understanding conflicting book reviews has implications far beyond books, ranging from opinions on merchandise, electronic devices, information services, to opinions on wars, religious, and environmental issues. Advances in this area have the potential of making substantial contributions to the assessment of the underlying credibility of evidence, the strength of arguments, diverse perspectives, and expectations. Choosing this topic has distinct advantages: no prior domain knowledge required, easy to interpret and evaluate results, potentially extensible applications to other genres.

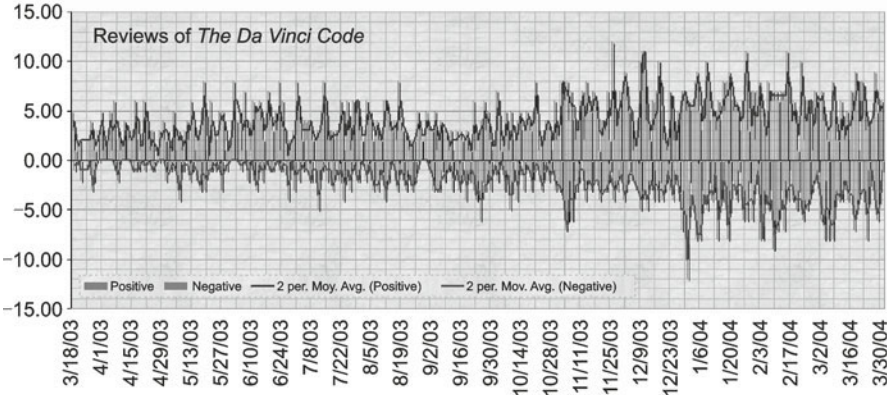


Fig. 7.1 The distribution of customer reviews of *The Da Vinci Code* on Amazon.com within the first year of its publication (March 18, 2003 ∼ March 30, 2004). Although positive reviews consistently outnumbered negative ones, arguments and reasons behind these reviews are not apparent. Source: (Chen, Ibekwe-SanJuan, SanJuan, & Weaver, 2006).

The Da Vinci Code is controversial. It attracted many positive reviews and negative reviews. What made it a bestseller? Which aspects of the book were favorably reviewed? Which aspects were criticized? More generally, will we be able to apply the same technique to other bestsellers, movies, cars, electronic devices, innovations, and scientific work? Ultimately, what are the reasons and turning points behind a success, a failure, a controversial issue, or conflicting information from multiple perspectives?

Sentiment analysis is a closely related topic, which aims to identify underlying viewpoints based on sentimental expressions in texts. Pang and Lee (2004) presented a good example of classifying movie reviews based on sentiment expressions. They used text-categorization techniques to identify sentimental orientations in a movie review and formulated the problem as finding minimum cuts in graphs. In contrast to previous document-level polarity classification, their approach focuses on context and sentence-level subjectivity detection. The central idea is to determine whether two sentences are coherent in terms of subjectivity. It is also possible to locate key sentimental sentences in movie reviews based on strongly indicative adjectives, such as *outstanding* for a positive review or *terrible* for a negative review. However, such heuristics should be used with considerable caution because there is a danger of overemphasizing the surface value of such cues out of context.

The majority of relevant research is built on the assumption that desirable patterns are prominent. Although this is a reasonable assumption for patterns associated with mainstream themes, there are situations in which such assumptions are not viable, for example, detecting rare and even one-time events and differentiating opinions based on their merits rather than the volume of voice.

7.1.2 Terminology Variation

Terminology variation is concerned with symbolic relations between terms and how they can be related through several types of variations and transformations (Daille, 2003). Variations of terms refer to the transformation of a term to a conceptually related term through linguistic operations such as morphological, syntactic, and semantic operations. Fig. 7.2 illustrates five types of transformations.

The TermWatch system provides functions for terminology variation studies, especially for extracting terms, identifying term variation relations, and clustering terms (Ibekwe-SanJuan, 1998; Ibekwe-SanJuan & SanJuan, 2004). TermWatch utilizes LTPOS[1] for term extraction and uses a hierarchical clustering algorithm called *Classification by Preferential Clustered Link* (CPCL) to identify and group terminological variations.

The CPCL algorithm first groups conceptually related terms together. A group of terms are conceptually related if they share a common head word. They may have different modifiers as in the examples of *ingenious plot* and *clever plot*. A group of terms are linguistically related if they can be transformed from one to another using one of the terminology variation operations

[1] http://www.cogsci.ed.ac.uk/~mikheev/tagger_demo.html

modifier

head word

Operations	Term	Term Variation
Syntactic (adding a modifier)	secret society	ancient secret society
Syntactic (adding a head word)	clever plot	clever plot twist
Syntactic (changing a modifier)	renowned Harvard professor	famous Harvard professor
Syntactic (changing a head word)	secret book	secret agenda
Semantic (synonymous)	ingenious plot	clever plot

Fig. 7.2 Linguistic operations underlying term variations.

such as spelling variants, WordNet semantic variants, and modifier variations. The algorithm then iteratively clusters groups of terms based on relations representing a considerable change. For example, the change from *secrete book* to *secrete agenda* is regarded as substantial. The procedure consists of several

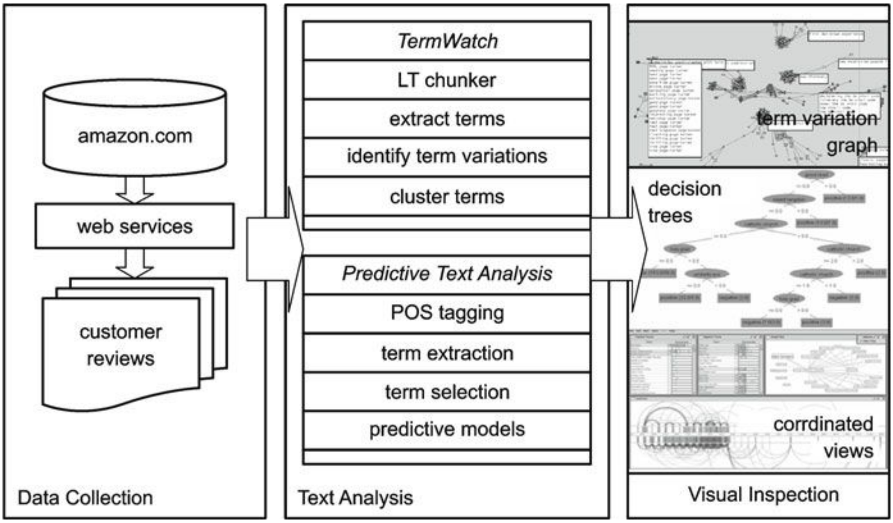


Fig. 7.3 The overall structure of our approach.

steps: data collection, term variation analysis, time series visualization of term variants, classification based on selected terms, and content analysis (see Fig. 7.3).

7.1.3 Reviews of The Da Vinci Code

Customer reviews of *The Da Vinci Code* were retrieved from Amazon.com using Amazon's web services (AWS). Amazon customer reviews are based on a 5-star rating system. 5 stars are the best and 1 star is the worst. In our study, reviews with 4 or 5 stars are regarded as positive reviews. Reviews with 1 or 2 stars are recoded as negative. Reviews with 3 stars are not used in this analysis.

As shown in Table 7.2, during the timeframe of the analysis, the number of positive reviews was about as twice as negative reviews. The average length of positive reviews is approximately 150 words and 9 sentences, whereas negative reviews are slightly longer, 200 words and 11 sentences on average. These reviews are generally comparable to news and abstracts of scientific papers in terms of their length.

Table 7.2 Statistics of the Corpus.

Corpus	Reviews	# Chars (mean)	#Words (mean)	#Sentences (mean)
Positive	2,092	1,500,707 (717.36)	322,616 (154.21)	19,740 (9.44)
Negative	1,076	1,042,696 (969.05)	221,910 (206.24)	12,767 (11.87)
Total	3,168	2,543,403	544,526	32,507

Our goal is to verify the feasibility of predicting the positions of reviews with a small set of selected terms. In addition, we expect decision trees to serve as an intuitive visual representation for analysts to explore and understand the role of selected terms as specific evidence in differentiating conflicting opinions. If we use selected terms to construct a decision tree and use the positive and negative categories of reviews as leaf nodes, the most influential terms would appear towards the root of the tree. We would be able to explore various alternative paths to reach positive or negative reviews.

In order to put the predictive power of our decision trees in context, we generate additional predictive models of the same data with other widely used classifiers, namely the naïve Bayesian classifier and support vector machine (SVM) classifier. We expect that although decision trees may not give us the highest prediction accuracy, it should be a worthwhile trade-off given the interpretability gain.

We use the procedure as follows. Reviews are first processed by part-of-speech tagging. Noun phrases are extracted with stopwords removed and the last word of each term stemmed. We include adjective as part of the phrases to capture emotional and sentimental expressions. Log likelihood tests are then used to select terms that are not purely high frequent, but influential in differentiating reviews from different categories. Selected terms represent an aggressive dimensionality reduction, ranging from 94.5%∼99.5%. Selected terms are used for decision tree learning and classification tests with other classifiers.

SVM can be used to visualize reviews of different categories. Each review is represented as a point in a high-dimensional space $\boldsymbol{S}$, which contains three independent subspaces $\boldsymbol{S}_p, \boldsymbol{S}_q$, and $\boldsymbol{S}_c : \boldsymbol{S} = \boldsymbol{S}_p \oplus \boldsymbol{S}_q \oplus \boldsymbol{S}_{c.}\boldsymbol{S}_p$ represents a review purely by positive reviews. Similarly, $\boldsymbol{S}_q$ represents a review in negative review terms only and $\boldsymbol{S}_c$ represents reviews with both positive and negative review terms. In other words, a review is decomposed into three components to reflect the presence of positive review terms, negative review terms, and terms that are common in both categories. Note that if a review does not contain any of these selected terms, then it will not have a meaningful presence in this space. All such reviews are mapped to the origin of the high-dimensional space and they are excluded from subsequent analysis.

The optimal configuration of the SVM classifier is determined by a number of parameters, which are in turn determined based on a k-fold cross-validation (Chang & Lin, 2001). This process is known as model selection. A simple grid search heuristic is used to find the optimal parameters in terms of the average accuracy so as to avoid the potential overfitting problem.

Table 7.3 shows the statistics of the term extraction and clustering by TermWatch. We describe these results in more detail in the following sections.

Table 7.3 Multi-layered feature selection using TermWatch.

Review categories	Terms	Classes	Components	Unique features
Positive	20,078	1,017	1,983	879
Negative	14,464	906	1,995	2,018

Fig. 7.4 helps to identify common characteristics of positive reviews of the book. For example, many reviewers found the book a page turner, with a wide variety of minor variations, such as an amazing page turner or an episodically page turner. It indicates that the popularity of the book is in part due to its gripping plots. The ability to group these terms together is a distinct advantage for reducing the complexity of the entire terminology.

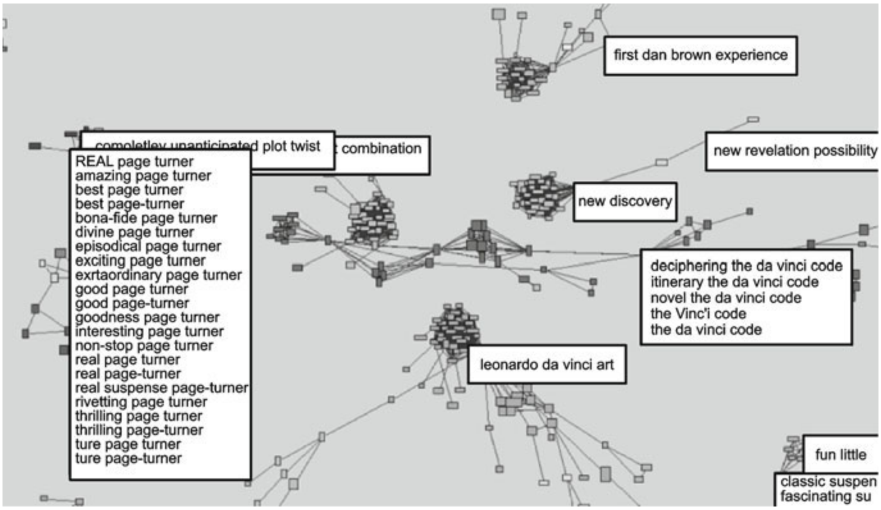


Fig. 7.4 Terms extracted from positive reviews are clustered based on both syntactic and semantic relationships. Source: (Chen et al., 2006).

7.1.4 Major Themes

A term variation network has three levels: clusters are shown at the highest level, then components, and finally terms at the lowest level.

7.1.4.1 Positive Reviews

The largest cluster labeled *leonardo da vinci art* in the network of terms associated with positive reviews is surrounded by the clusters *literary fiction*, *the complete dead sea scroll*, *harvard professor*, and *isaac newton*. The structure of this cluster is highly interconnected and its content appears to be coherent as it captures the main facets of the positive reviews: comments on the major characters (*Prof Langdon*), the praises (*great storytelling, clever story, gripping novel, historic fiction*), other major characters (*Sophie Neveu, Leonardo Da Vinci, Sir Isaac Newton*).

Another main cluster *Da Vinci code fuss* is also about the book itself (*the da vinci code fuss, the da vinci novel, the da vinci code review*). They were grouped into the same cluster because of the terminological variation (here modifier substitution).

The *Da Vinci code fuss* cluster is linked to another cluster labeled *the vinci code*, which in turn connects to another cluster labeled as *mary magdelene legend*. The *mary magdalene* cluster is concerned with the historical plausibility of events, people and organizations described in the book. For instance, there is much controversy about the supposed liaison between Mary Magdalene and Jesus Christ. Other much debated topics are the roles of the Prieure de Sion and Opus Dei organizations, the effects of the historical events as depicted in the book on religious faith of today's Christians, the research the author claimed to have carried out to back up his version of the historical events. Because of the varied nature of the terms in this cluster, most of the links are due to associations (co-occurrence).

An isolated sub-network deals with the author's writing history: his next, previous or new books. Apparently, the terminology used to talk about this in the reviews is distinct from the terms used to praise the current book, hence the isolation.

7.1.4.2 Negative Themes

Terms that appeared most frequently in negative reviews include *mary magdalene*, *opus dei*, *the holy grail*, *too much*, *art history*, *good book*, *page turner*, *secret society*, *the last supper*, *conspiracy theory*, and *villain*. The negative reviews questioned the historical and religious foundations of the book which the author (Dan Brown) presented as "truth based on research." The author's claims came under ferocious criticisms by the negative reviewers who undertook to prove point by point that the author is misleading his reader. The most controversial point is centered on the religious events portrayed in the book such as the supposed love affair and subsequent marriage between Jesus Christ and Mary Magdalene. Indeed, the term *mary magdalene* is consistently featured in all negative reviews from the first year since the book was published in March 2003.

7.1.5 Predictive Text Analysis

The predictive text analysis of the book reviews serves two objectives: to validate the predictive power of selected terms and to provide a visual representation for analysts to explore and understand the role of these terms in reviews of different categories.

Terms are ranked differently by document frequency and log-likelihood ratio. As shown in Table 7.4, terms with high document frequency tend to be descriptive of the book being reviewed (e.g., *book*, *story*, *novel*, *fiction*), whereas terms with a high log-likelihood ratio tend to be more related to opinions, judgments, and recommendations (e.g., *money*, *hype*, *great read*, *disappoint*, *waste*).

Table 7.5 summarizes the number of terms selected by log-likelihood values and the accuracies of three classifiers with 10-fold cross-validation. The original set of extracted terms contains 28,763 terms. The dimensionality reduction rates range from 94% to 99.5%. In contrast, if we select terms based on their document frequencies (>= 2), there will be 6,881 terms and the accuracy of classification with a C4.5 decision tree is 68.89%, which is below all the models with log likelihood tests. More importantly, decision trees (C4.5) are relatively stable in terms of 10-fold cross-validation accuracies (slightly over 70%), whereas SVM models have more than 80% of accuracy, which means the selected terms are good candidates to categorize these reviews. These classifiers are available in data mining software *Weka* (Witten & Frank, 1999).

Table 7.4 Terms ranked differently by document frequency (DF) and log-likelihood ratio.

Term	DF	Log-likelihood	Term	DF	Log-likelihood
book	2456	2.99	money	83	68.27
story	697	14.25	write	179	66.61
reader	571	0.23	hype	146	61.37
character	561	59.32	character	561	59.32
da vinci code	559	10.85	author	504	53.04
novel	539	0.00	great read	92	48.40
fiction	536	4.89	couldn't	135	48.10
time	512	21.17	disappoint	39	46.77
author	504	53.04	waste	33	39.26
plot	499	17.25	don't waste	22	37.63

Table 7.5 Classification accuracy on 10-fold cross-validation.

Log Likelihood (p-level)	Selected Terms	C4.5	NaiveBayes	SVM
0.05	1,666	70.26	77.54	84.59
0.01	360	71.67	76.67	83.14
0.001	146	70.01	75.74	81.72
Doc Freq (>=2)	6,881	68.89		

Although using document frequencies as feature selection metrics may give comparable results to metrics such as information gain, it is not as efficient as other metrics if aggressive dimensionality reduction is desired (Yang & Pedersen, 1997). Terms selected by log-likelihood tests of the presence and absence of a term in relation to the category of a review are visualized in Fig. 7.5 with GGobi[2]. It shows the majority of terms have relatively low document frequencies. On the other hand, terms such as, money, hype, character, and great read are selected with quite different document frequencies.

7.1.5.1 Decision Trees

Two decision trees are shown in Figures 7.6 and 7.7 to illustrate how they may facilitate tasks for differentiating conflicting opinions. The top of the tree contains terms that strongly predict the category of a review, whereas terms located in lower part of the tree are relatively weaker predictors.

In the 2003 decision tree the presence of term *great read* predicts a positive review (Fig. 7.6). Interestingly, if a review does not mention *great read*, *Robert*

[2] http://www.ggobi.org/

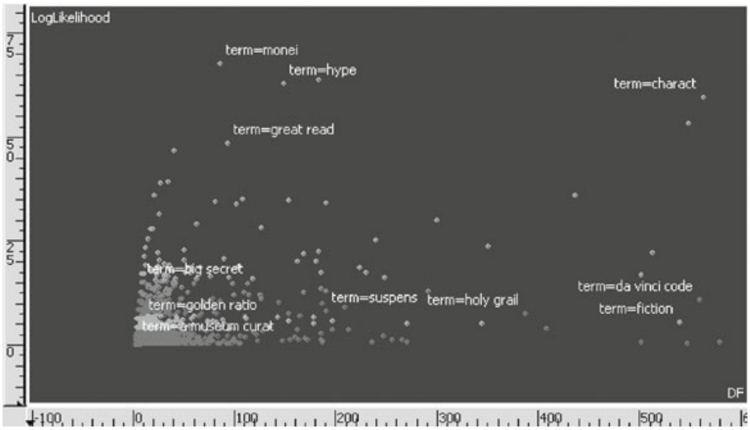


Fig. 7.5 Distributions of selected terms. The colors of dots indicate the statistical significance level of the corresponding terms, namely green (< 0.001), blue (p=0.001), red (=0.01), and pink(=0.5). Source: (Chen et al., 2006). (see color figure at the end of this book)

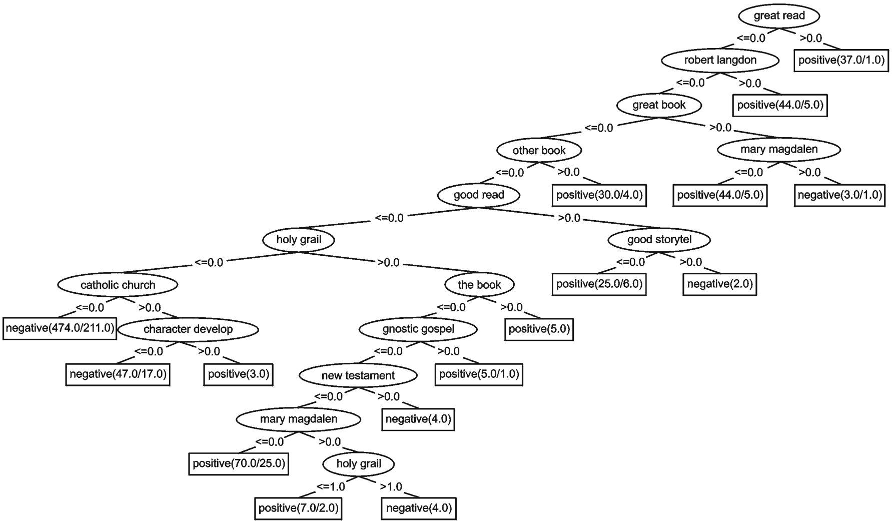


Fig. 7.6 A decision tree representation of terms that are likely to differentiate positive reviews from negative reviews made in 2003. Source: (Chen et al., 2006).

Langdon, but talks about *mary magdalen*, it is more likely to be a negative review. Similarly, the branch at the lower right corner shows if a review mentions both *mary magdalen* and *holy grail*, then is also likely to be a negative review. In comparison, in the 2004 decision tree the term *first page* predicts a positive review (see Fig. 7.7). If a review mentions both *mary magdalen* and *holy grail*, then it is likely to be a negative review.

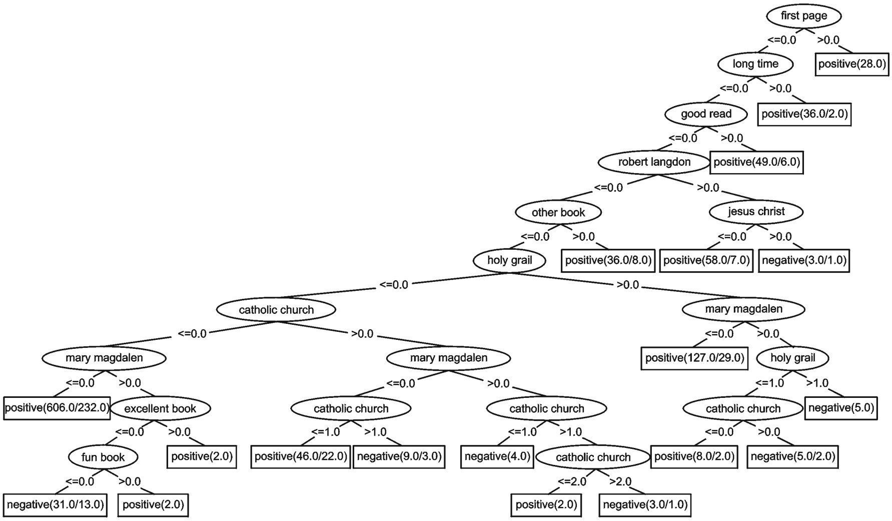


Fig. 7.7 A decision tree based on reviews made in 2004. Source: (Chen et al., 2006).

7.1.5.2 Classifying Reviews by Active Terms

Linguistically active terms refer to terms that have many variants. Active terms are used to label the clusters and they represent a much smaller portion of the phrases, which is 8.3% of the noun phrases extracted by the LT-chunker in LTPOS.

Since these active terms have not been selected based on their occurrence in the reviews, they are not expected to be the best candidates for indexing reviews in a classification task. We index reviews with TermWatch cluster labels if terms in a cluster appear in reviews. These terms are expected to be closely related to cluster labels as they are linked by a short chain of variations. We then generated an additional decision tree using 60% of the data as a training set. The resultant decision tree correctly classifies 68% of the remaining reviews. This accuracy is lower than those obtained by previous classifiers, but it remains interesting since it is mainly based on long multiword terms, which tend to have much lower frequencies than single-word terms.

The accuracy of this decision tree relies on the 30 most active terms, i.e. they have the greatest number of variants in reviews. For example, *robert langdon story* has 250 variants in reviews of which 85% are positive. Similarly, terms such as *opus dei website*, *millennium-old secret society* and *historical fact revelation* have more than 100 variants in reviews of which 66% are positive.

Browsing terms not included in the decision tree model is also informative. For example, each of the terms *anti christian*, *secret grail society blah blah*

blah, and *catholic conspiracy* has only 6 variants in reviews, all negative, identifying readers shocked by the book.

Browsing the interrelationship between reviews and TermWatch clusters reveals topics that appear in both categories positive/negative and thus ignored by decision trees. As it turns out, each of the terms like *jesus christ wife*, *mary magdalene gospel*, *conspiracy theory* and *christian history* have more than 50 variants that are almost evenly distributed between positive and negative reviews.

The perspective of term variation helps to identify the major themes of positive and negative reviews. For negative reviews, the heavy religious controversies raised by the book are signified by a set of persistent and variation rich terms such as *mary madgalena, opus dei,* and*the holy grail*, and none of these terms ever reached the same status in positive reviews. Much of the enthusiasm in positive reviews can be explained by the perspective that the book is a work of fiction rather than scholarly work with discriminating terms such as *vacation read, beach read,* and *summer read.*

Fig. 7.8 show an opinion differentiation tree regarding the product of video iPod. This decision tree model's accuracy of classification is as high as 91.49%. The presence of the term *video quality* predicts a positive review, whereas *battery life* is a sign for a negative review. The more specific term *short battery life* appears in the tree at the 6th level from the top.

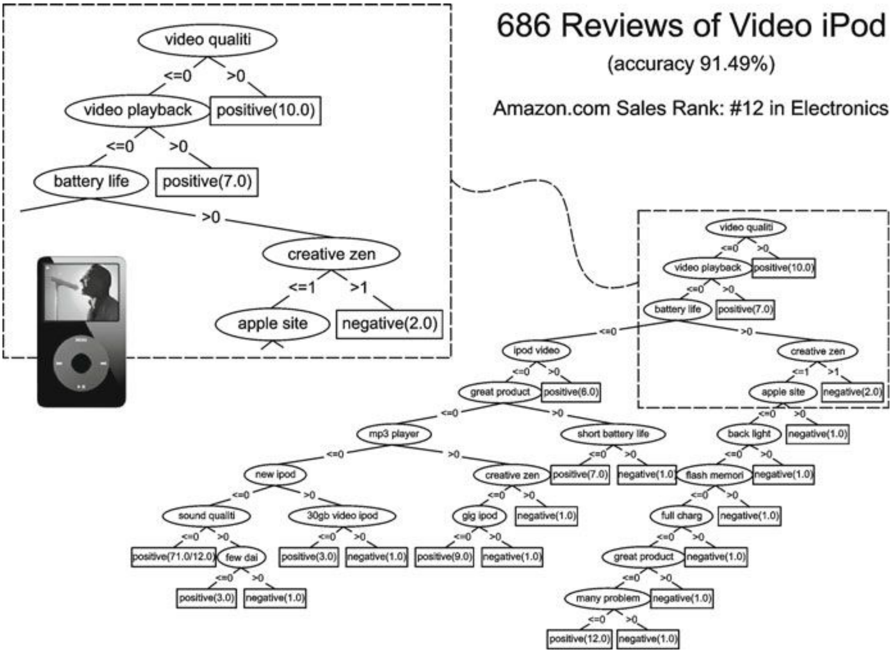


Fig. 7.8 An opinion differentiation tree of Video iPod reviews.

The same technique is applicable to identifying emerging topics in a field of study. Fig. 7.9 depicts a decision tree representation of noun phrases that would identify a new topic or an old topic in terrorism research. As shown in the decision tree, the term *terrorist attack* is an old topic with respect to the timeframe of 2004−2005. In contrast, *mental health* is a new topic. The *risk assessment* of *biological weapon* in particular is a new topic (at the time of data analysis).

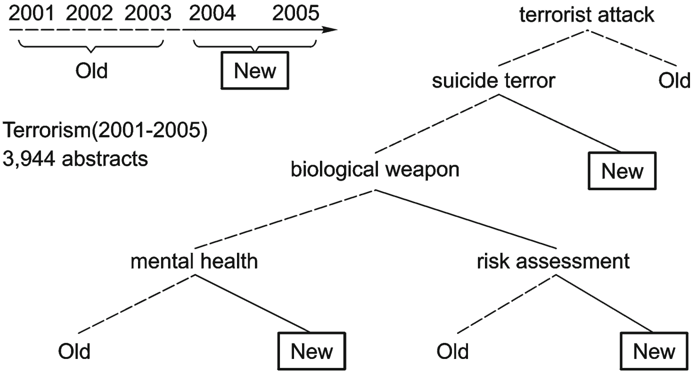


Fig. 7.9 Representing emerging topics in 3,944 abstracts of publications on terrorism.

In summary, the analysis of the reviews of the *Da Vinci Code* illustrates the nature of conflicting opinions. One of the primary reasons people held different opinions is because they view the same phenomenon through different perspectives. The examples of video iPod and terrorism research illustrate the potential of the approach for a wider range of applications.

7.2 Analyzing Unstructured Text

In our analysis of customer reviews, a useful source of evaluative information is the ratings provided by customers. In general, such rating information is not always readily available. How do we make a sense of a pile of documents in unstructured text? In this section, we will introduce a new approach to identify underlying concepts and relationships between concepts from unstructured text. What makes this approach unique is that there is no assumption of the availability of any prior knowledge of the domain in question. In connection to the terminology variation perspective, we contrast what is changing and what is not in natural language passages so that we can identify natural groupings of expressions that correspond to some latent concepts.

7.2.1 Text Analysis

Several techniques are becoming increasingly mature to facilitate tasks of understanding a large amount of text. At the document level, information visualization techniques can present a global view of an entire document collection based on term frequency and other statistical models of information (Chalmers, 1992; Havre, Hetzler, Whitney, & Nowell, 2002; Hetzler, Whitney, Martucci, & Thomas, 1998; Kohonen, 1995). Users are able to explore and examine various groupings of text and drill down to individual documents. At the word and sentence levels, visualizations have also been made to depict associations of words (Callon, Courtial, Turner, & Bauin, 1983; Rip & Courtial, 1984; Tijssen & Vanraan, 1989). Visualizations at all these levels can reveal insightful patterns and have found many valuable applications. On the other hand, there remains a considerable gap between the two major levels of text visualization that prevents users from moving back and forth smoothly.

The relatively overlooked middle ground is in part due to the lack of text visualization that explicitly addresses the core concepts and assertions in an input source of text. Although one may access the original text via a network visualization of co-occurring words to trace the exact sentences in which a specific word occurs, the gist of the assertion is not readily accessible from the visualization. Consequently, the analyst still faces the challenge of identifying and synthesizing key assertions and statements made in the source text.

We introduce a new method of text visualization to bridge the gap between document-centric and word-centric approaches. Unlike text visualization approaches such as Phrase Nets (Ham, Wattenberg, & Viégas, 2009) and Word Tree (Wattenberg & Viégas, 2008), we focus on supporting text navigation over aggregations of assertions rather than tracing text word by word. In this sense, the new method provides an intermediate layer of interactive visualization between document-focused and word-focused visualizations. Thus, the new method should facilitate the transitions between existing layers of visualizations.

The new method utilizes natural language processing techniques for part-of-speech tagging and rules defined by regular expressions for pattern matching. Two major types of patterns are designed to capture concepts and predicates from the source text. We use the term predicate broadly in this article, including subject, verb, and object. Extracted patterns are organized in a tree structure and visualized as a DOI tree, a degree of interest tree (Budiu, Pirolli, & Fleetwood, 2006; Card & Nation, 2002), using the prefuse (Heer, 2007) implementation at the backend. For each source of text, two trees are generated: a concept tree and a predicate tree. Salient branches in these trees reveal salient concepts and assertions. Predicates are cascaded in the tree construction process, which may lead to a long trail of interlocked predicates. The new method enables the user to compare two sources of text by adding patterns found in one source to the patterns found in another. We will discuss the details shortly.

Fig. 7.10 The structure of a concept tree. Each sub-tree corresponds to an underlying concept.

Text summarization provides techniques to construct a short summary of multiple documents. A good summary should have a sufficient and balanced coverage with minimal redundant information (Sparck Jones, 1999). Automatic multi-document summarization typically constructs a short summary by selecting the most representative sentences from a set of topically related documents. Teufel and Moens (2002) have proposed an intriguing strategy for summarizing scientific articles based on the rhetorical status of statements in an article. Their strategy specifically focuses on identifying the new contribution of a source article and its connections to earlier work. Automatic summarization techniques have been applied to areas such as identifying drug interventions from MEDLINE (Fiszman, Demner-Fushman, Kilicoglu, & Rindflesch, 2009). However, much of text summarization is still a black-box approach. Users have neither control nor access to the underlying selection mechanisms. Statistically-driven summarization algorithms may miss important sentences.

Text visualization is one of the longest running streams of research in information visualization. Earlier systems include Bead (Chalmers, 1992), ThemeRiver (Havre et al., 2002), IN-SPIRE (PNNL), and various Self-Organized Maps (Kohonen, 1995). Many of the earlier systems focus on the big picture and on groupings of words and documents.

A different line of text visualization focuses more on preserving the sequential order of the original text in visualized format. For example, TextArc (Paley, 2002) visualizes words but also provides a way to guide users to follow the original text. Ben Fry created overviews of different editions of Darwin's book *The Origin of Species* so that one can see what changes were made between these editions (Fry, 2009).

Recent work such as Word Tree and Phrase Nets, both available from ManyEyes (Viégas, Wattenberg, Ham, Kriss, & McKeon, 2007), provides a new way to read sentences in text. In Word Tree, one can enter a word and start to trace multiple treads of sentences anchored on the common words. In Phrase Nets, one can select a specific type of relation from a predefined set to group words into a network. We can easily tell that the most frequently used words in IEEE InfoVis abstracts (2000–2009) are *data*, *visualization*, and *information*. However, users may not easily identify how exactly these words are connected and in what contexts.

It is worth noting that Phrase Nets intentionally chose to avoid using complex natural language parsers. Their results are remarkable considering the simplicity of their design. In this chapter, we want to explore the strengths and weaknesses of combining pos-tagging (Toutanova, Klein, Manning, & Singer, 2003) and regular expression-based pattern matching in identifying deeper semantic patterns in unstructured text.

The notion of degree of interest (DOI) is introduced by George Furnas in his work on fisheye views (Furnas, 1986). The visualization of DOI trees is revisited by Heer and Card (Heer & Card, 2004). DOI trees have unique advantages that are particularly appropriate for displaying deeper semantic patterns in natural language forms because they provide an affordance for the user to read from a node to its children nodes. The children nodes provide the immediately accessible context of the parent node. By using simple mouse over techniques, one can leverage the extra layer of the interface between the user and the original text so that entries in the tree can be partial sentences or stems and rely on the interaction to flush out the rest of the sentences.

Another advantage of the DOI tree layout, as implemented by prefuse, is the emergent edge bundling effect, which becomes evident at the global level. Such effects provide important clues for the analyst to zoom in.

7.2.2 Searching for Missing Links

Don Swanson developed a strategy that generates new hypotheses by linking previously disparate bodies of knowledge, for example, between fish oil and Raynaud's syndrome (Swanson, 1986). Our own theory of transformative discovery (Chen, Chen, Horowitz, Hou, Liu, & Pellegrino, 2009) is built on a generalized structural hole theory (Burt, 2004) and it states that many transformative discoveries are made by connecting two or more previously isolated substructures of knowledge. For example, one of the most fundamental discoveries in string theory in physics was the discovery that two systems previously thought to be different were proven to be mathematically equivalent. Another example is in terrorism research, where it was previously believed that only people who have direct traumatic experience could develop symptoms of post traumatic stress disorder, but it was later proven that people who were not at the scene could also develop the disorder due to a vivid coverage of traumatic events by mass media (Chen, 2006).

In order to identify such candidate hypotheses, it becomes necessary for an analyst to have an efficient access to various assertions and claims embedded in otherwise unstructured text. As the first step to achieve this goal, the analyst needs to be able to transcend from reading the source text sequentially. Since he/she may deal with an unfamiliar source of text, the analytical process needs to start with an overview of all patterns rather than a predefined query.

Research in citation-based trend analysis provides additional motivations to the work. A typical way to analyze emerging trends in a scientific domain is to analyze the structure and dynamics of its literature by forming a network of references and then studying the network dynamics (Chen, 2006). Analyzers often run clustering algorithms to divide the document space into clusters of documents. Interpreting the nature of such clusters has been a bottleneck of the analytical process. What the analyst needs at this stage are patterns that characterize precise relations expressed in terms of hypotheses and findings, which are not readily captured by statistical methods. In other words, patterns must reflect natural language expressions.

Analysts often need to differentiate two instances of text, either two documents on related topics, or two samples of text from different time points. Historians of science, for example, need to study the variations of terms in order to establish reasons why a scientific theory was rejected before and why it was accepted later on. In the case of the continental draft theory, its acceptance relies on much fundamental conceptual changes (Thagard, 1992).

7.2.3 Concept Trees and Predicate Trees

Analysts need to answer some common questions in text analysis. What are the core concepts in the source text? What are the contexts in which they appear? What are the major attributes of a concept? What are the most common assertions in the source text? What are the longest chains of cascading predicates? What are the differences between two sources of text in terms of concepts and predicates?

The basic design consists of two organizational and visualization components — concept trees and predicate trees. Both are hierarchical representations of natural relations between several types of words found in the original sources of unstructured text. A concept tree represents the groupings of words surrounding nouns. A predicate tree represents the accumulated and emergent structure after all the predicates found by the pattern matching regular expressions in the input text are merged. For simplicity, the term predicate in this design includes the subject and the predicate (i.e. the verb and objective nouns).

We are motivated to address one of the most fundamental challenges for processing unstructured text — the lack of structure. The idea is to construct a meaningful organizing structure out of the unstructured text such that the organizing structure can server an intermediate role between browsing high-level global structures such as a document space and understanding lower-level local details such as the relationships among various types of words and the formation of sentences in a given input source. Our basic assumption is that the processing procedure has no access to additional semantic resources except the input data per se. The reason for making such a restricted assumption is that we want to establish a baseline for the further development of such processing procedures and we also want to identify how far the existing natural language processing and general-purpose programming techniques can achieve the goal.

Our approach is inspired by an observation that is intrinsically related to the notion of the degree of interest (DOI). According to the commonly known explanation of DOI, variations of perceived details in a scene are the function of the viewer's interest. The function reflects where the viewer's interest is placed. Usually, we pay more attention to things next to us. In contrast, we may pay less and less attention to things further and further away from us. If we turn this thinking to natural languages, we recognize something strikingly similar — variations of descriptive details in text are the function of the writer's interest. If we are writing about a few topics, we tend to use naturally more words to describe, clarify, differentiate, and iterate topics that we think are more important than the rest of them. We tend to find more examples and take into account more perspectives than otherwise. As a result, the more important topics will be surrounded by far richer varieties of words than the rest of topics. The design of the new procedure draws upon this observation and focuses on identifying the core of such a concentration in text as the symbol of a concept. In addition, we decide to concentrate on the predicate chain of the subject, the verb, and the object in a sentence so as to simplify the original sentence and make it easy for the user or the analyst to decide whether there is sufficient interest to pursue. We expect that given a sufficiently large amount of input, it becomes more likely that the basic structure of a sentence is to be found more and more frequently, especially for important topics. As such instances accumulate, emergent patterns may appear. Such patterns are expected to be insightful for making sense of unstructured text. They may play instrumental roles in facilitating the further visual analysis of seemingly unstructured text.

7.2.3.1 Procedure

The flow chart in Fig. 7.11 illustrates the key components of the procedure and their communications.

The procedure starts with the selection of sources of text. The user may select a single document, multiple documents, and a directory of documents as the initial input data source. The selected documents as a whole are referred as a source of the procedure. The current prototype supports two sources. The text in the selected source is subsequently processed by part-of-speech (POS) tagging. The result is that each and every word in the original text is annotated with a POS tag. For example, the noun tree is tagged as tree/*nn* and the verb run is tagged as grow/*vb*.

The next step, pattern matching, is to identify segments of word sequences based on their POS tags and then organize various parts of such segments according to heuristics on implied hierarchical relations. For example, the

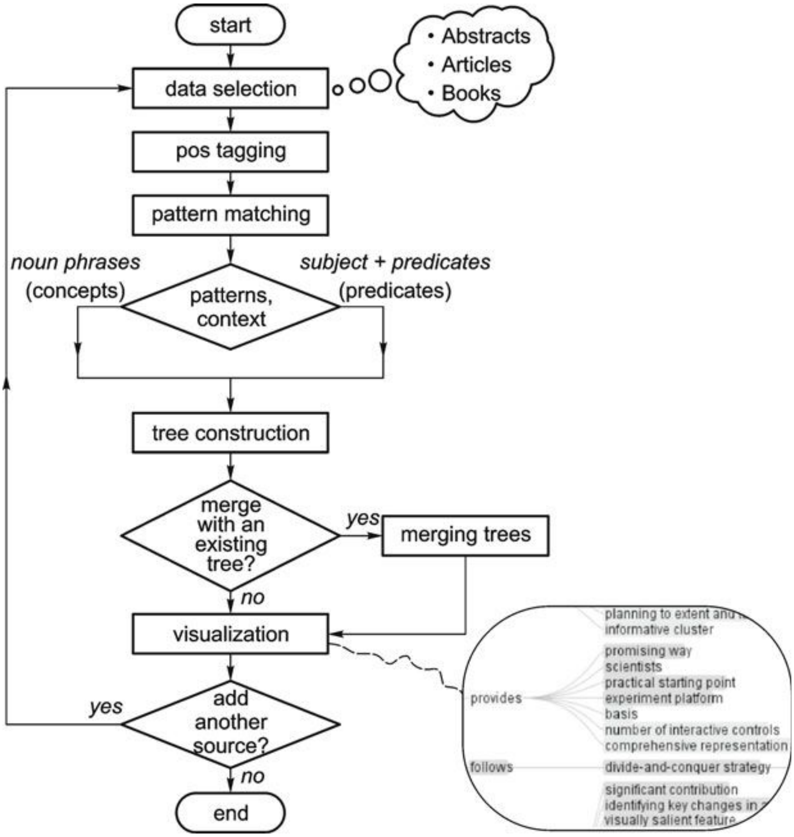


Fig. 7.11 The flow chart of the procedure of the method.

noun phrase *large-scale network* can be split into two parts *large-scale* and *network*. The noun *network* is known as the head noun. The *large-scale* part is known as the modifier of the head noun. Thus the word network represents a concept and it will be stored as a parent node in a hierarchical representation. The term *large-scale* can be seen as an attribute of the concept and it will be stored as the child node to the parent network. The pattern matching process is illustrated in Fig. 7.12.

Table 7.6 summarizes the major patterns defined by regular expressions over POS-tagged text. To make the construction of these patterns easier, we use a bottom-up approach. Starting with the basic building blocks such as nouns, verbs, and adjectives, more complex patterns are built by joining these building blocks. For example, the predicate pattern is defined in terms of subject, verb, and object, which are in turn defined in terms of noun phrases and verb groups.

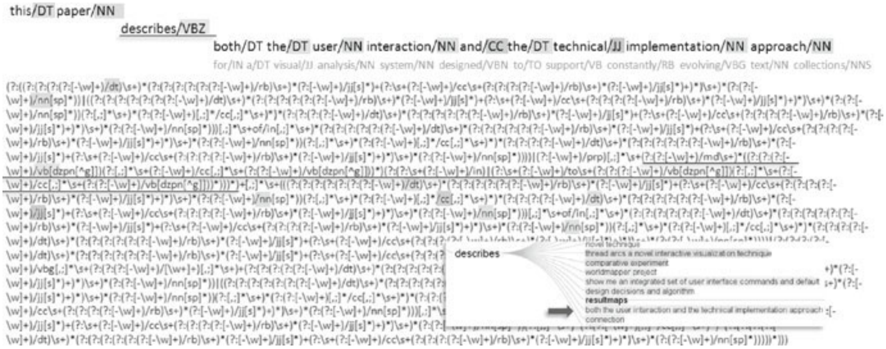

this/DT paper/NN
describes/VBZ
both/DT the/DT user/NN interaction/NN and/CC the/DT technical/JJ implementation/NN approach/NN
for/IN a/DT visual/JJ analysis/NN system/NN designed/VBN to/TO support/VB constantly/RB evolving/VBG text/NN collections/NNS
describes
novel technique
thread arcs a novel interactive visualization technique
comparative experiment
worldmapper project
show me an integrated set of user interface commands and default
design decisions and algorithm
resultmaps
both the user interaction and the technical implementation approach
connection

Fig. 7.12 The regular expression of the subject-predicate pattern consists of 3,480 characters. The sentence on the top of the figure is tagged first. Corresponding patterns are matched based on rules defined in the regular expression for predicates. Identified patterns are added to the tree.

Table 7.6 Building blocks of patterns.

Pattern	Definition	Example	Length(Pattern)
noun	(article OR adjective)* + word /*nn*[sp]*	heavy and cold rain	181 characters
noun phrase	various combinations of nouns and other types of words	Information visualization research	858 characters
subject	noun OR noun phrase OR word/*prp*	this article	1,057 characters
simple verb	word /*vb*[dzpn[ˆg]]	discover	27 characters
verb	Various combinations of verbs	could have been discovered	257 characters
concept	noun phrase, including noun of noun	exotic plant	858 characters
predicate	subject + verb + (noun phrase OR noun OR word/*vbg*)	we + introduce + a new algorithm	3,480 characters

The length of a regular expression pattern can serve as an interesting benchmark. A future improvement of the method may shorten the length while maintaining the overall coverage and accuracy of the extraction and pattern matching step. Furthermore, more complex and lengthy patterns may be designed to capture more subtle and complex assertions in unstructured text.

A concept tree and a predicate tree will be generated as the results of the pattern matching step. The concept tree consists of all extended noun phrases found in the text. Phrases with the same head noun are aligned up first, then their attribute words. For example, *heterogeneous information space* and *exploration of an information space* share the concept of information space. Both *heterogeneous* and *exploration* are organized as the children nodes of information space as its attributes. Similarly, if a predicate of *we* + *introduce* + *a new algorithm* is found, the construction algorithm will first check if there is already a node we in the predicate tree. If the we node exists, then check if it has a child node named *introduce*, then finally check the next level for a possible place for the term *new algorithm*. Leading a's, an's, and the's are omitted from the tree representations.

Different trees of the same type can be merged by the same rules. One may also incorporate more advanced graph mining techniques to find identical substructures in such trees. Furthermore, a predicate tree can be refined and enhanced by normalizing its subjective nouns and objective nouns with concept nodes in a concept tree. For example, the predicate *we* + *introduce* + *a new algorithm* can be standardized as *we* + *introduce* + *algorithm*.

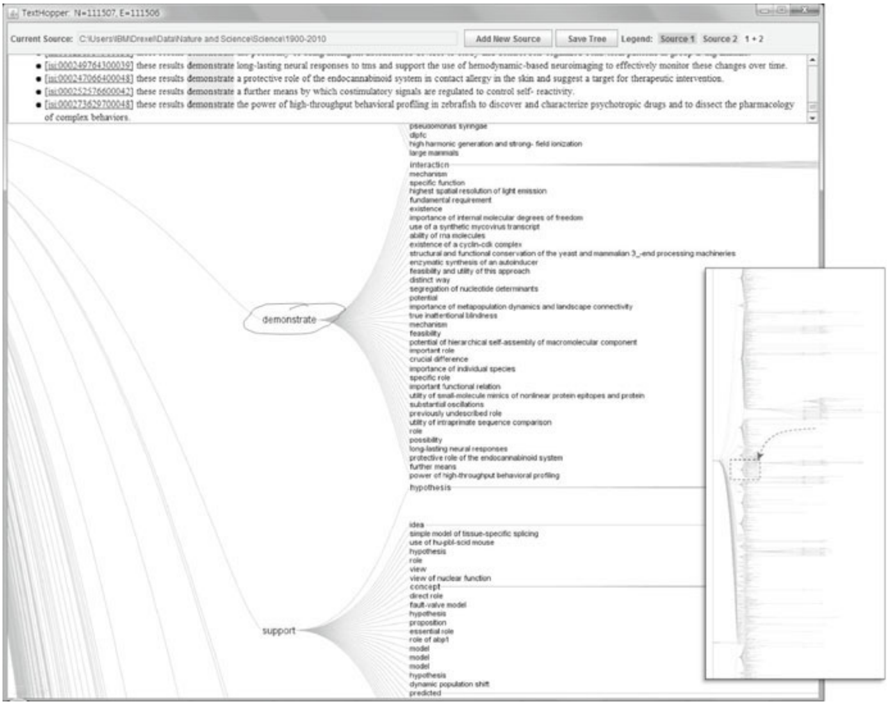
The merge step integrates tree representations generated from multiple sources but retains the identification of each of the original sources so that one can compare the contributions from different sources side by side. Merging patterns from different sources follows the same rules for within-source constructions, except that the origin source of each pattern is now taken into account. Internally, patterns from the first source and those from the second source can be distinguished. Three colors are used to encode the three possible relationships: pink — two patterns are from the same first source, green — both are from the same second source, and yellow — they have different sources. In principle, one may choose to allow the addition of a series of sources, although it may considerably increase the cognitive burden of the analyst to recognize patterns from too many possible sources.

The final step is for users to interact with visualizations and explore the underlying data through the hierarchical representation of concepts and predicates. Using tree representations can leverage many existing tree visualization tools. For example, prefuse supports a number of ways to visualize a tree structure, including the balloon layout, the radial layout, and the node-and-link layout. We experimented different layouts and finally decided that the node-and-link layout provides the most appropriate design option for us.

The prefuse implementation of the DOITree provides a robust vehicle for the baseline prototype. It allows the user to zoom in and out easily. It scales reasonably well. The response rates decrease for concept trees with more than 200,000 nodes. On the other hand, pruning trees may become advisable because it is likely to increase the clarity and the speed of interaction.

The analyst is able to access the original context of each instance associated with a node by moving the mouse cursor over the node. The contextual information shows the names of the sources and the complete sentences in which a pattern is found.

The user interface of a prototype is shown in Fig. 7.13. It shows a predicate tree that was constructed from 110-year *Science* articles' abstracts (1900-2010). The predicate tree consists of 111,507 nodes, including both nouns and verbs that form a predicate. The circled term *demonstrate* is part of the predicate pattern: *these results demonstrate*. The full sentences in context


TextHopper: N=111507, E=111506
Current Source: C:\Users\IBM\Drexel\Data\Nature and Science\Science\1900-2010
Add New Source
Save Tree
Legend: Source 1 Source 2 1 + 2

- [isi:000249764300039] these results demonstrate long-lasting neural responses to tms and support the use of hemodynamic-based neuroimaging to effectively monitor these changes over time.
- [isi:000247066400048] these results demonstrate a protective role of the endocannabinoid system in contact allergy in the skin and suggest a target for therapeutic intervention.
- [isi:000252576600042] these results demonstrate a further means by which costimulatory signals are regulated to control self- reactivity.
- [isi:000273629700048] these results demonstrate the power of high-throughput behavioral profiling in zebrafish to discover and characterize psychotropic drugs and to dissect the pharmacology of complex behaviors.

pseudomonas syringae
dipfc
high harmonic generation and strong- field ionization
large mammals
interaction
mechanism
specific function
highest spatial resolution of light emission
fundamental requirement
existence
importance of internal molecular degrees of freedom
use of a synthetic mycovirus transcript
ability of rna molecules
existence of a cyclin-cdk complex
structural and functional conservation of the yeast and mammalian 3_-end processing machineries
enzymatic synthesis of an autoinducer
feasibility and utility of this approach
distinct way
segregation of nucleotide determinants
potential
importance of metapopulation dynamics and landscape connectivity
true inattentional blindness
demonstrate
mechanism
feasibility
potential of hierarchical self-assembly of macromolecular component
important role
crucial difference
importance of individual species
specific role
important functional relation
utility of small-molecule mimics of nonlinear protein epitopes and protein
substantial oscillations
previously undescribed role
utility of intraprimate sequence comparison
role
possibility
long-lasting neural responses
protective role of the endocannabinoid system
further means
power of high-throughput behavioral profiling
hypothesis
idea
simple model of tissue-specific splicing
use of hu-pbl-scid mouse
hypothesis
role
view
view of nuclear function
concept
direct role
fault-valve model
hypothesis
support
proposition
essential role
role of abp1
model
model
model
hypothesis
dynamic population shift
predicted

Fig. 7.13 The user interface of a prototype. The portion of the predicate tree shown in the figure represents the pattern of "these results demonstrate ...", which forms a small branch of a 111,507-node hierarchy constructed based on 110-year *Science* article abstracts (1900 – 2010).

are displayed on the top of the screen and instances of the current focal node in the tree are highlighted in yellow.

7.2.3.2 Examples of Use

We tested the prototype on four types of text, namely abstracts of scientific papers, abstracts of patents, full-length articles, and monographs. In particular, abstracts of scientific papers include IEEE InfoVis papers (2000 – 2009), *Science* (1900 – 2010). Abstracts of patents include 823 *Google* patents and 15,227 *Yahoo*! patents. Full-length articles include Shneiderman's highly cited *The Eyes Have It* (Shneiderman, 1996), and two journal papers by Chen on *CiteSpace* (Chen, 2004; Chen, 2006). The prototype was applied to two books: one is the *Brokerage and Closure* by Burt (Burt, 2005) and the other is Charles Darwin's 1872 edition of *The Origin of Species* (6th ed.) (Darwin, 1872), excluding the glossaries and index of the book. We intend to use this set of examples to set a baseline for subsequent evaluations.

We recorded the total number of sentences and words found in each text, the percentage of nouns and verbs identified by part-of-speech tagging, and the sizes of both the concept tree and the predicate tree. We are particularly interested in the relationship between the average length of sentences and the overall coverage rate (the percentage of sentences that are found with concept and predicate patterns). We also recorded the runtime taken to completion. The experiment used an IBM ThinkPad T500 with duo processors of 2.53GHz and 3GB of RAM and the Java Runtime version of 1.6.0_11. The results are shown in Tables 7.7 and 7.8.

Table 7.7 Datasets tested.

Source	Type	Sentences	Words	Nouns (%)	Verbs (%)
Yahoo Patents(15227)	Abstract	9,342	189,662	40.72	13.84
Google Patents (823)	Abstract	4,372	83,586	39.40	14.40
InfoVis (2000 – 2009)	Abstract	2,139	42,472	34.40	12.35
Darwin (1872)*	Book	5635	197,332	23.05	13.77
Burt (2005)	Book	5,201	112,556	30.94	12.87
Science (1900 – 2000)	Abstract	98,370	2,062,010	37.09	10.98
Chen(2004)	Article	358	6440	30.45	13.23
Shneiderman(1996)	Article	258	4,500	34.04	12.51
Chen (2006)	Article	659	10,831	33.60	12.55

*Excluding glossaries and the index.

Table 7.8 Records are sorted by the coverage rate (% sentences in trees).

Source	Type	Concepts	Predicates	Coverage (% sentences)	Runtime (mill. Sec)
Yahoo Patents(15227)	Abstract	15,227	12,082	67.90	71,666
Google Patents (823)	Abstract	7,091	5,393	63.63	577
InfoVis (2000 – 2009)	Abstract	5,364	2,535	58.85	6,957
Darwin (1872)*	Book	13,583	5,538	50.22	279,622
Burt (2005)	Book	11,187	5,896	49.51	147,826
Science (1900 – 2000)	Abstract	279,932	111,506	49.44	4,852
Chen(2004)	Article	899	375	41.06	8,722
Shneiderman (1996)	Article	747	279	37.60	6,514
Chen (2006)	Article	1,463	589	34.90	14,756

Fig. 7.14 suggests that the average length of sentences in text may be correlated with the overall coverage rate. Patent abstracts and InfoVis abstracts have particularly higher coverage rates than other sample datasets tested. It

seems to suggest a general trend of the higher word-per-sentence rate, the higher the coverage with the three exceptions in the middle. A possible explanation is that all the ones in the middle are primarily information science and computer science related. In contrast, the *Science* abstract dataset is more in line with others perhaps because of its multidisciplinary nature and that the performance of the pos-tagger may be deteriorated for biomedical literature. Yahoo and Google patent abstracts yielded the highest coverage rate, followed InfoVis abstracts. The two books are placed in the middle of the coverage range. Individual full-length articles have the lowest coverage rate among all the sample datasets.

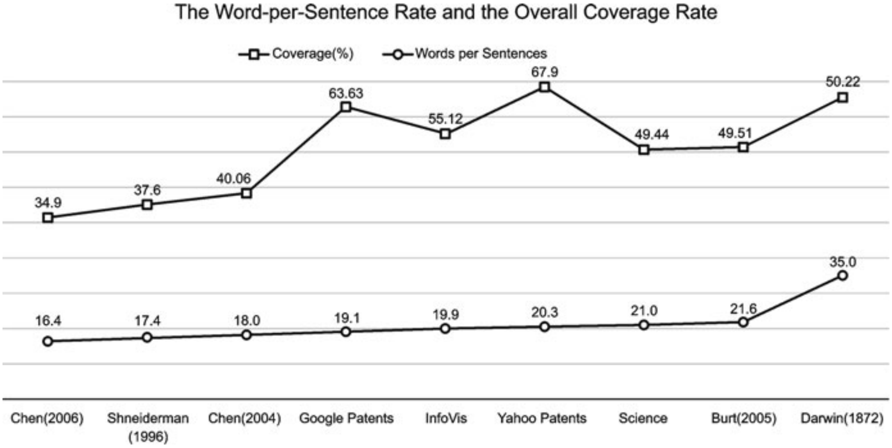


Fig. 7.14 Words from longer sentences may be more likely to be included in the trees.

7.2.3.3 Scenarios of Use

Three major scenarios of use are outlined as follows, although one can certainly use the method in many ways.

The first scenario is the analysis of a multi-document collection. The analyst needs to identify the major topics and claims along with their original contexts. The analyst may have a few options. At a macroscopic level, he/she may choose to visualize inter-document relations based on similarity measures derived from term frequencies and decompose the entire collection into topically related clusters of documents. At a microscopic level, she may also use tools such as Phrase Nets and Word Tree to develop an understanding of the content at word and sentence levels. Between the two levels, she may explore the concept tree and the predicate tree to obtain the information that would enable her to make better use of the tools operating at the other two levels. For example, as shown in Fig. 7.15, the major topics of InfoVis are shown as the nodes with many children nodes in its concept tree, such as *data*, *information*, and *visualization*. Furthermore, the analyst can zoom in and out to obtain more contextual details of a concept, for example, all sorts of adjectives and nouns that were found surrounding the term *data* in this particular source of text.

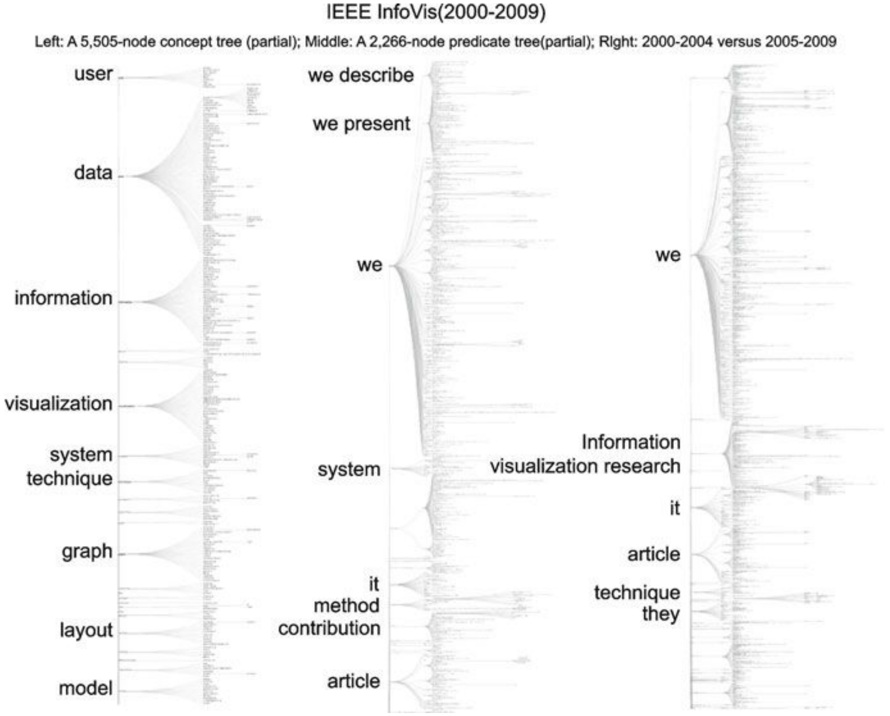


Fig. 7.15 A concept tree, a predicate tree, and a comparative predicate tree of the InfoVis abstract dataset.

One potential advantage of using a tree representation over alternatives such as a network is the improved clarity and readability. Major topics are visually prominent in tree representations without risking potential problems of distractive cross links that are often seen in network representations of such datasets. Another potentially significant advantage is that the analyst may quickly track down the most dominating rhetorical patterns in the data source. Combining with the knowledge of what major topics are, the analyst can browse through the entire collection without being restricted to the boundaries of individual documents; in fact, the grouping patterns allow the analyst to compare similar assertions across different documents in the collection. In the current prototype, one can find all the instances of a pattern by moving the mouse mover the node in the tree representation.

The second scenario is the comparison of two sources or two samples from the same source. If we want to know how InfoVis papers in the earlier years differ from InfoVis papers in the later years, we can split the 10-year dataset into two parts, generate the trees for the first part, and add the secondary part to these trees. Fig. 7.16 is an example of such merged trees. Patterns found in the first sample are shown in pink; those found in the second sample shown in green; and patterns in common are in yellow. These predicates are associated with the subject node *article*. We can see that the most commonly used patterns in the InfoVis dataset include *article* + *presents* + *, *article* + *introduces* + *, *article* + *proposes* + *, and *article* + *describes* + *. These rhetorical statements are probably also common for the majority scholarly publications. Scrolling deeper down the tree reveals more semantically focused statements, for example, *GPU* + *requires* + *data parallel programming*.

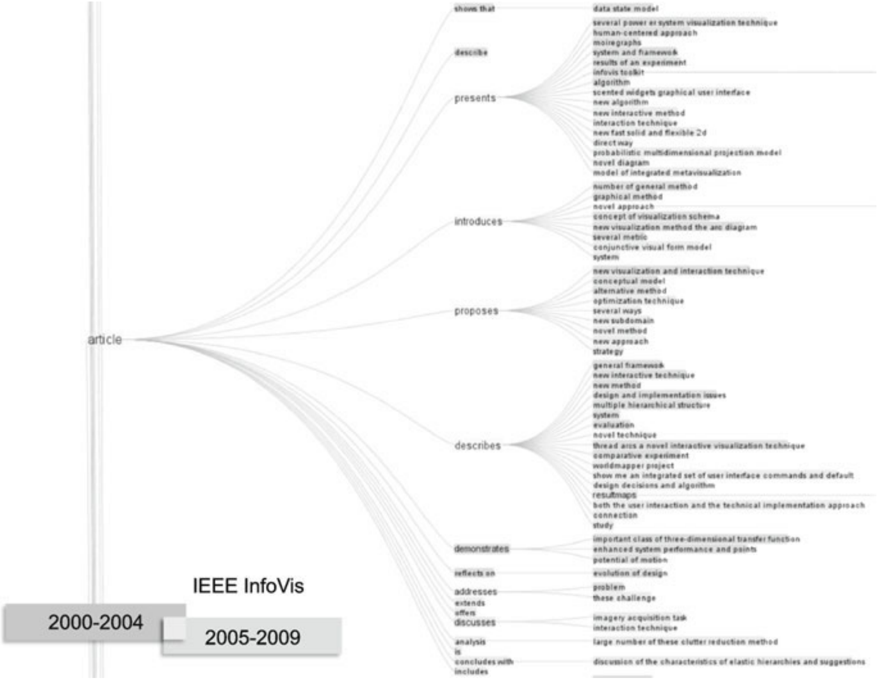

article
shows that
data state model
describe
presents
several power er system visualization technique
human-centered approach
moiregraphs
system and framewok
results of an experiment
infovis toolkit
algorithm
scented widgets graphical user interface
new algorithm
new interactive method
interaction technique
new fast solid and flexible 2d
direct way
probabilistic multidimensional projection model
novel diagram
model of integrated metavisualization
introduces
number of general method
graphical method
novel approach
concept of visualization schema
new visualization method the arc diagram
several metric
conjunctive visual form model
system
proposes
new visualization and interaction technique
conceptual model
alternative method
optimization technique
several ways
new subdomain
novel method
new approach
strategy
describes
general framework
new interactive technique
new method
design and implementation issues
multiple hierarchical structure
system
evaluation
novel technique
thread arcs a novel interactive visualization technique
comparative experiment
worldmapper project
show me an integrated set of user interface commands and default
design decisions and algorithm
resultmaps
both the user interaction and the technical implementation approach
connection
study
demonstrates
important class of three-dimensional transfer function
enhanced system performance and paints
potential of motion
reflects on
evolution of design
addresses
problem
these challenge
extends
offers
discusses
imagery acquisition task
interaction technique
analysis
large number of these clutter reduction method
is
concludes with
discussion of the characteristics of elastic hierarchies and suggestions
includes
IEEE InfoVis
2000-2004
2005-2009

Fig. 7.16 A merged predicate tree from IEEE InfoVis papers in two periods of time: 2000 – 2004 and 2005 – 2009.

In addition to collection-collection comparisons, one may need to compare two documents, or one document against one collection. The notion of source makes it flexible. One can instantiate a source with one document or multiple documents. This flexibility allows an analyst to apply standard clustering algorithms on a collection of text and then apply the new approach introduced here to each cluster and compare different structures by treating clusters as incoming text sources. The new approach provides an alternative to make sense individual clusters by using techniques such as text summarization. The advantage of the visual approach over traditional text summarization is that the analyst can move back and forth smoothly across different levels of words, sentences, documents, and the entire data set, which would reduce the cognitive burden of the analyst.

The third scenario is the study of a lengthy document. In this scenario, the analyst can use the interactive visualization of the concepts and predicates as an indexing mechanism. Since all the instances and contexts of a concept can be found within the subtree of the concept, it becomes easy for the analyst to access them. We illustrate this scenario with two book examples.

Brokerage and Closure is written by sociologist Ronald S. Burt (2005) to introduce the concepts of brokerage and closure in social networks and their practical implications. Fig. 7.17 shows the top of the concept tree and the top of the predicate tree, where prominent patterns are usually positioned. For example, it is visually evident that the book has many ways to describe *people*, *network*, *trust*, and *ideas* as shown on the left concept tree. Similarly, the predicate tree on the right reveals that the leading actors in the book are *you*, *we*, *they*, and *it*.

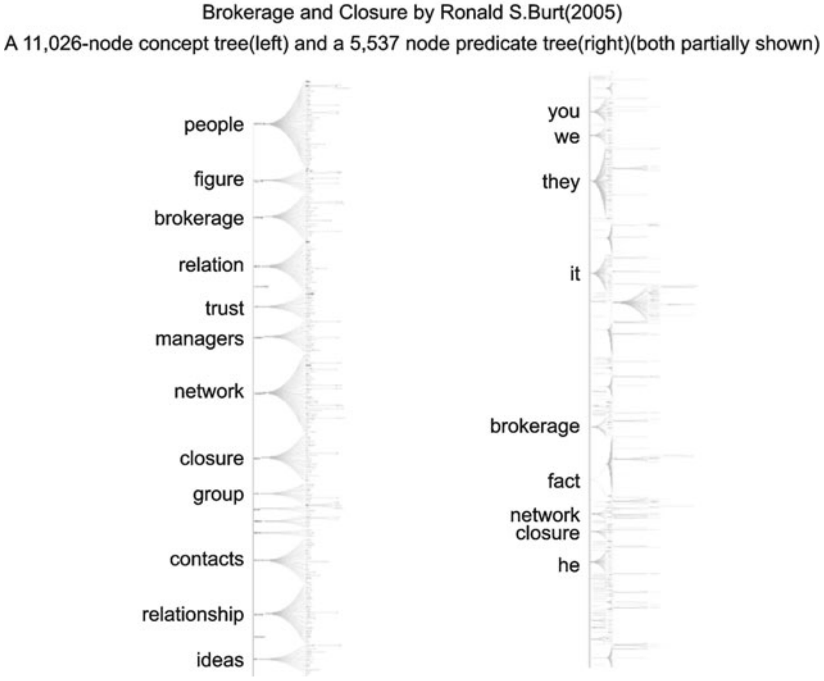

Brokerage and Closure by Ronald S.Burt(2005)
A 11,026-node concept tree(left) and a 5,537 node predicate tree(right)(both partially shown)
people
figure
brokerage
relation
trust
managers
network
closure
group
contacts
relationship
ideas
you
we
they
it
brokerage
fact
network
closure
he

Fig. 7.17 A concept tree and a predicate tree of Ronald Burt's 2005 book.

The example in Fig. 7.18 is Darwin's classic *The Origin of Species* (6th ed.). Predictably, *species* is the most prominent concept, followed by *forms*, *varieties*, *differences*, *animals*, *plants*, and *group*. Its predicate tree includes patterns such as *forms of life* + *are* + *, *it* + * + *, *we* + * + *, *they* + * + *, and *many exotic plants* + *have* + *.

7.2.3.4 Further Improvement

Design and technical issues are discussed as follows, which could serve as the basis for further improvement.

The highest coverage rate among the 9 datasets is the Yahoo patent abstracts (67.90%). The lowest coverage rate is a full-length article published in the Journal of the American Society for Information Science and Technology (JASIST) (34.90%). The InfoVis abstract dataset has the 3rd highest coverage (58.85%). The current set of heuristics and regular expression patterns obviously cannot process all the sentences. Some of the problems are due to the POS-tagging errors. Some are due to the complexity of sentence structures.

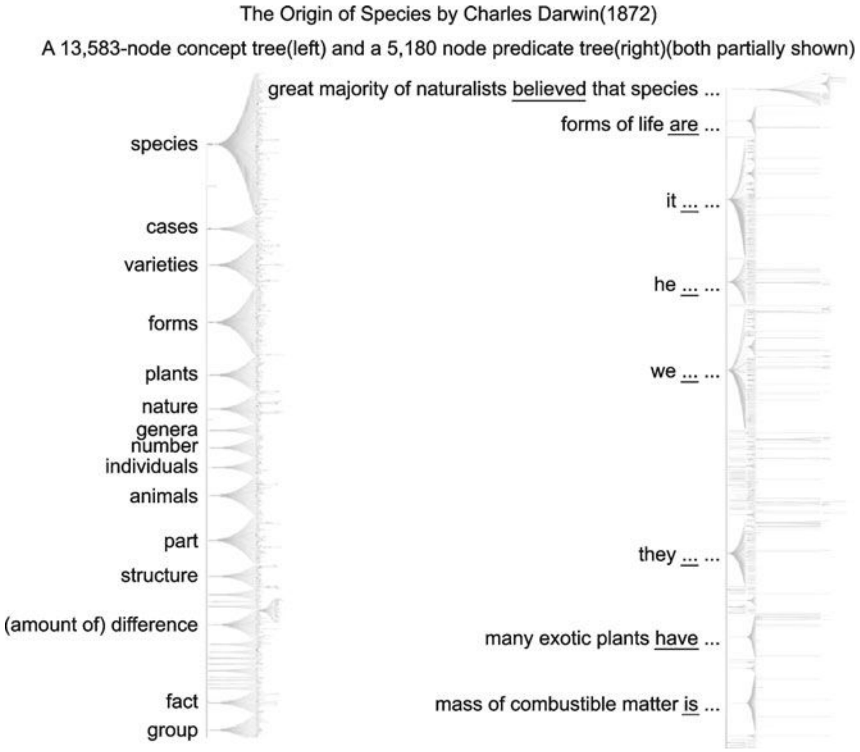

The Origin of Species by Charles Darwin(1872)
A 13,583-node concept tree(left) and a 5,180 node predicate tree(right)(both partially shown)
species
cases
varieties
forms
plants
nature
genera
number
individuals
animals
part
structure
(amount of) difference
fact
group
great majority of naturalists believed that species ...
forms of life are ...
it ... ...
he ... ...
we ... ...
they ... ...
many exotic plants have ...
mass of combustible matter is ...

Fig. 7.18 A concept tree and a predicate tree of Darwin's *the Origin of Species.*

We list some examples below and hope they can serve as test data for further enriched and enhanced heuristics and pattern matching rules. The challenges for the further improvement include improving the efficiency of the pattern matching mechanisms by shortening the lengths of regular expressions, expanding the coverage of the processing procedure so that a wider variety of sentences can be handled, and improving the time efficiency of the algorithm by shortening the overall runtime.

1) Our/PRP$ tool/NN replaces/VBZ manual/JJ and/CC in/IN some/DT cases/NNS pen-and-paper/JJ based/VBN analysis/NN tasks,/NN and/ CC we/PRP discuss/VBP how/WRB user/NN feedback/NN was/VBD incorporated/VBN into/IN iterative/JJ design/NN refinements/NNS
2) The/DT method/NN is/VBZ widely/RB used/VBN in/IN many/JJ research/NN fields/NNS including/VBG biology,/JJ geography,/NN statistics,/NN and/CC data/NN mining/NN
3) However,/NNP most/RBS dendrograms/NNS do/VBP not/RB scale/VB up/RP well,/JJ particularly/RB with/IN respect/NN to/TO problems/ NNS of/IN graphical/JJ and/CC cognitive/JJ information/NN overload/ NN
4) The/DT overview/NN displays/VBZ only/RB a/DT user-controlled,/JJ limited/JJ number/NN of/IN nodes/NNS that/WDT represent/VBP the/DT “skeleton”/NN of/IN a/DT hierarchy/NN
5) The/DT contribution/NN of/IN the/DT paper/NN includes/VBZ a/DT new/JJ metric/JJ to/TO measure/VB the/DT “importance”/NN of/IN nodes/NNS in/IN a/DT dendrogram;/NN the/DT method/NN to/TO construct/VB the/DT concise/NN overview/NN dendrogram/NN from/ IN the/DT dynamically-identified,/JJ important/JJ nodes;/NN and/CC measure/NN for/IN evaluating/VBG the/DT data/NNS abstraction/NN quality/NN for/IN dendrograms/NNS
6) We/PRP evaluate/VBP and/CC compare/VBP the/DT proposed/VBN method/NN to/TO some/DT related/JJ existing/VBG methods,/NN and/CC demonstrating/VBG how/WRB the/DT proposed/VBN method/NN can/MD help/VB users/NNS find/VB interesting/JJ patterns/NNS through/IN a/DT case/NN study/NN on/IN county-level/JJ U.S/NNP

There are more fundamental questions to be addressed. For example, what is the nature of these extracted patterns? Are they more rhetorical than semantic? The top-level predicate patterns appear to be rhetorical such as *we* + *propose* + *a new algorithm* .... Since we are interested in facts and hypotheses, we are looking for predicates such as *smoking* + *causes* + *lung cancer*. We found such statements in the 111,506-node predicate tree of Science, but we do not have the statistics of how many of them and do not yet have reliable features that would allow us to distinguish scientific statements from rhetorical ones. Purifying predicate trees becomes necessary in order to build analytical reasoning functions. Another way to refine the structure of predicate trees is to utilize the structure of concept trees, as we mentioned briefly earlier.

The scalability of the new method is also an important issue for further research. The current prototype can process full-length books. It handles the 110-year *Science* abstract dataset surprisingly well, although interacting with the 111,506-node predicate tree and the 279,932-node concept tree experienced considerable delays. The burden on the visualization program is partially due to the built-in storage of contextual information for each node. Using an indexing mechanism to store the bulk of data outside the tree structures would improve the performance.

The implementation of the prototype uses our own hand-crafted regular expression patterns. More sophisticated patterns may be developed and tuned up with natural language processing tools such as GATE. It will be also useful to compare the coverage and other benchmark scores with a wider range of natural language processing resources, including various pos-taggers.

The new method has several application implications. For example, the tree construction procedure can be used to develop an alternative indexing and ranking algorithm by weighing on the size of subtrees of a node. One may also derive semantic metrics based on the positions of nodes on such trees and measure the degree of interest between two nodes. This indexing potential has the advantage of preserving the original context and providing an easy to access interface to all the instances in multiple documents.

The new method provides an extra layer of interface between text visualization and text so that one can focus on exploring aggregated patterns as the intermediate linkage between specific words and sentences and their broader context. This extra layer enables the analyst to move back and forth across the boundaries of documents and focus on the essence of the text.

The emergent structural patterns enable users to identify the areas to pursue from the global overview of the entire dataset. The new method allows the analyst to contrast and compare two sources of text at various levels of detail, for example, in the study of patents from competing corporations or publications from different schools of thoughts.

Future work should address the issues discussed above. In addition, constructing an authoritative ontology of the information visualization field for comprehensive experimental tests, incorporating ontology construction techniques, and conducting user evaluation and field studies are among the promising routes to proceed.

7.3 Detecting Abrupt Changes

Michael Jackson, the legendary king of the pop, died on June 25, 2009. The news of his death created a surge of search volume on the Internet that was too much for Google News to handle.[3] The volcanic peak was marked as B in Fig. 7.19.

This type of surge is also known as a burst. Identifying such bursts in a timely manner is the primary goal for burst detection. While it is reasonably straightforward to track down reasons behind this type of burst as well as their timing and duration, it can be much more complex and challenging in other situations. Is there a visible burst of positive or negative reviews of *The*

[3] http://articles.cnn.com/2009-06-26/tech/michael.jackson.internet_1_google-trends-search-results-michael-jackson?_s=PM:TECH

Da Vinci Code in Fig. 7.1? Was there a burst of particular terms in those reviews? In the literature of a scientific field, do we expect to see a burst of a topic as it becomes suddenly popular? If a field is experiencing a Kuhnian paradigm shift, or a conceptual revolution, would we be able to detect a burst of articles representing the new paradigm?

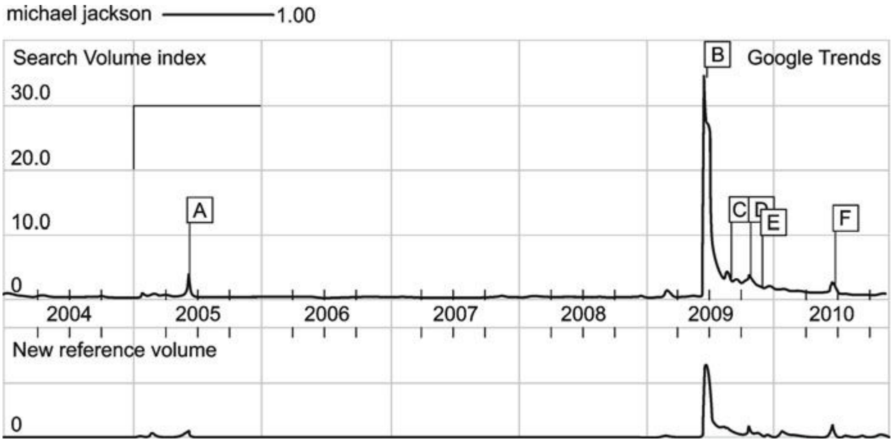


Fig. 7.19 The volcanic peak of search of "Michael Jackson." Source: Google Trends.[4]

7.3.1 A Burst of Citations

We have addressed some of these questions in earlier chapters in the context of structural analysis and the theory of discovery. In this chapter, we will show how burst detection can be used in comparative studies so that we would be able to answer questions such as whether positive reviews tend to burst before negative reviews, or when and how long the burst of a new paradigm would last.

First, let us look at an example of citation burst. Fig. 7.20 shows a part of a larger network of papers that were cited together in the literature of complex network analysis. This part is in fact the so-called largest connected component of the entire network. The node highlighted in the middle was Wasserman's representative work on social network analysis. The node is in a crucial position that connects two sub-networks of the component. It is a pivotal point in the development of the topic. The cluster above the pivotal node is labeled by the term *scale-free network*, the cluster below the pivotal node is labeled by the term *social network*. These two clusters corresponding to the earlier social network analysis that was mostly done by sociologists (the

[4] http://www.google.com/trends?q=michael+jackson

one below) and the more recent complex network analysis that was largely done by physicists.

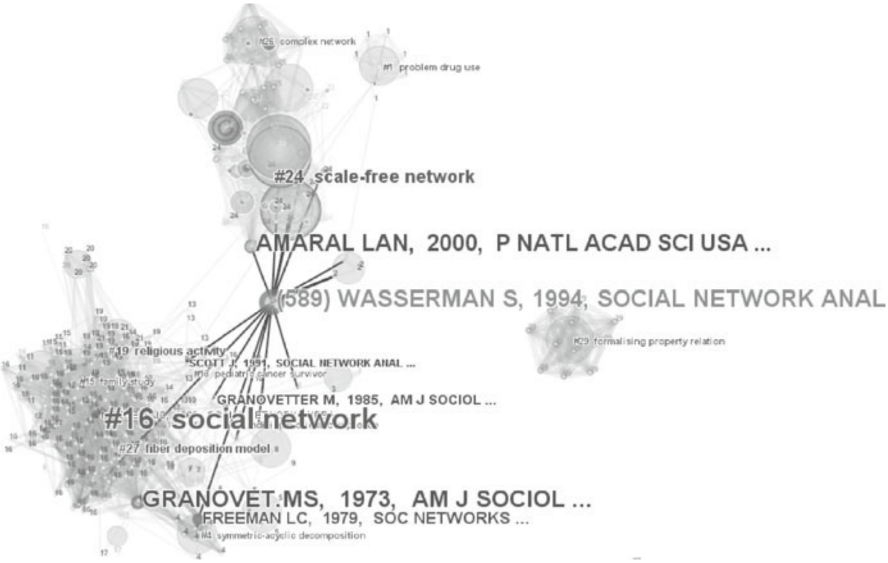

#25 complex network
#1 problem drug use
#24 scale-free network
AMARAL LAN, 2000, P NATL ACAD SCI USA ...
(589) WASSERMAN S, 1994, SOCIAL NETWORK ANAL
#29 formalising property relation
#19 religious activity
SCOTT J, 1991, SOCIAL NETWORK ANAL ...
#18 pediatric cancer survivor
#15 family study
GRANOVETTER M, 1985, AM J SOCIOL ...
#16 social network
#27 fiber deposition model
GRANOVET.MS, 1973, AM J SOCIOL ...
FREEMAN LC, 1979, SOC NETWORKS ...
#4 symmetric-acyclic decomposition

Fig. 7.20 The largest connected component of complex network analysis research (1980 – 2009).

The unique position of the pivotal node suggests that it is essentially one of the few that both communities had in common. Was there any burst of citation to the pivotal-point work? If we can detect a burst, what would it tell us about the evolution of the two communities?

As it turns out, there was indeed a burst of citation. The burst started in 1998 and lasted till 2000 (Fig. 7.21). Among other early citers in 1998, one of them was the groundbreaking paper for complex network analysis written by Watts and published in Nature. What is intriguing is the position of the burst in the entire 20-year history. The level of the citation counts during the burst period is not the highest on the hindsight, but the timing is much more meaningful for the purpose of identifying an emerging trend or a new field. The timing synchronized with the publication of the groundbreaking paper of the new complex network analysis. Furthermore, it was indeed cited by the groundbreaking paper. The citations of the pivotal work increased rapidly from 2002. The delay in part reflects how long it may take for knowledge diffusion and for the new paradigm to establish.

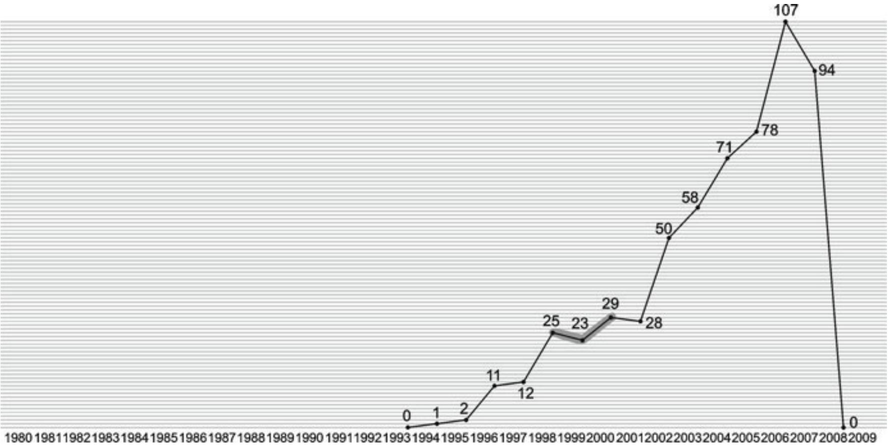


Fig. 7.21 A burst of citation was detected between 1998 and 2000.

7.3.2 Survival Analysis of Bursts

Survival analysis is a statistic method for answering questions concerning the time elapsed till the occurrence of an event, which is called survival time. Survival analysis looks at the probability of survival at various time points. Survival analysis also takes into account censored cases, which refer to cases that never experience the occurrence of an event throughout the observation time or drop-off the observation. A typical use of survival analysis is to compare the effectiveness of a new drug in comparison to that of an existing drug. For example, survival analysis can be used to assess whether a new painkiller is more effective than an existing painkiller. The survival time of a painkiller can be defined as the time elapsed from the time that the medicine is taken till the pain gets back. In this case, the returning pain is the event in question. If patients taking the new drug experience a higher survival probability than the existing drug, then the new drug is more effective.

We are dealing with scientific publications, citation patterns, grant proposals, and other types of knowledge representations. We are mainly concerned with a burst of citations to a published paper and a burst of the frequency of a topic. We can analyze the probability of survival at any time in terms of two types of events associated with a burst: (1) time elapsed prior to burst and (2) the length of the duration of a burst. Fig. 7.22 illustrates the two types of comparisons. Time to effect is the waiting time prior to the initial burst. Thus, the shorter the waiting time, the better it is. The same principle should hold for both citation bursts and topic bursts. Similarly, a longer duration of a citation burst is more preferable. So is a longer duration of a topical burst.

Fig. 7.23 illustrates a concrete example of the citation history of a highly

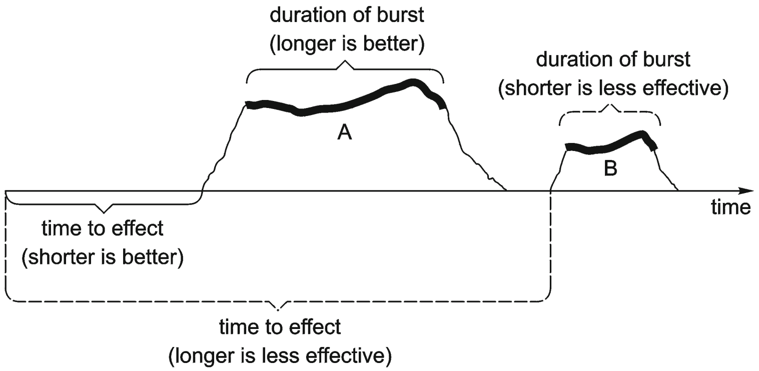


Fig. 7.22 Comparing time to effect and duration of burst with survival analysis.

cited paper in string theory written by Juan Maldacena. The paper was published in 1998. A citation burst was detected between 1999 and 2003. The waiting time was 1 year. A total of 288 cited references are shown in the network in Fig. 7.24. They are split into two groups by the median citation of 19.50, high- and low-citation groups, so that survival analysis can address the question whether papers in the two groups differ in terms of patterns associated with their citation bursts.

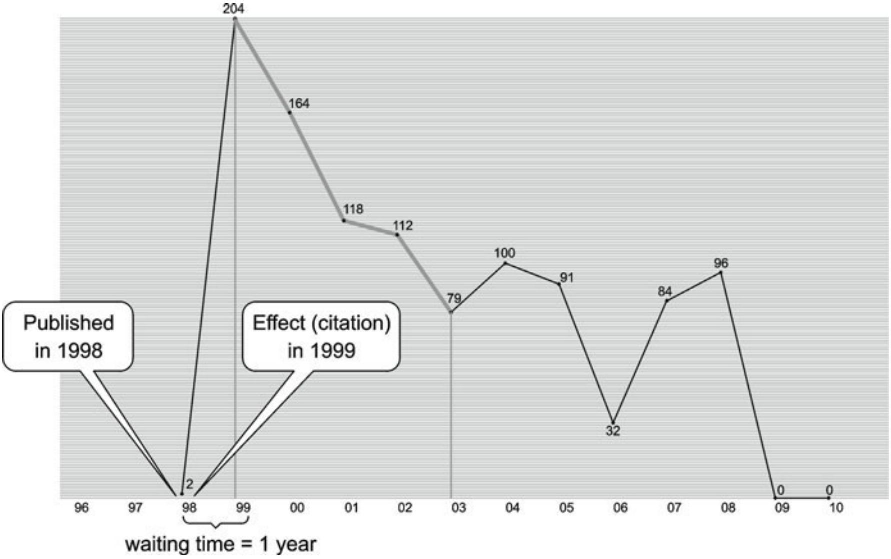


Fig. 7.23 The waiting time and the duration of a citation burst for a 1998 paper by Maldacena. The burst began in 1999, one year after its publication, and ended in 2003.

Survival analysis found that the highly cited reference group is likely to

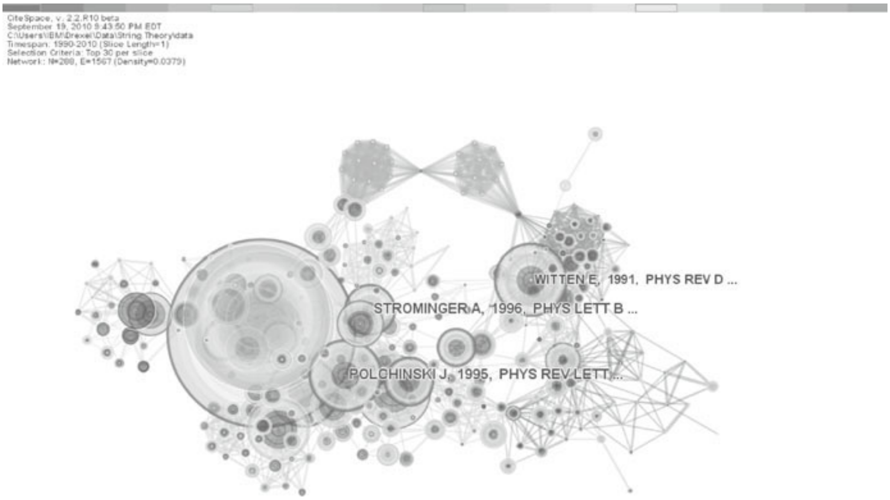

CiteSpace, v. 2.2.R10 beta
September 19, 2010 9:43:50 PM EDT
C:\Users\IBM\Drexel\Data\String Theory\data
Timespan: 1990-2010 (Slice Length=1)
Selection Criteria: Top 30 per slice
Network: N=288, E=1567 (Density=0.0379)
WITTEN E, 1991, PHYS REV D ...
STROMINGER A, 1996, PHYS LETT B ...
POLCHINSKI J, 1995, PHYS REV LETT ...

Fig. 7.24 The input data for survival analysis consists of 288 cited references as shown in this network. They are split into a high- and low-citation group by h-index.

have a citation burst sooner than the less cited papers. A highly cited paper on the average waits for about 2.2 years, whereas a less cited paper waits for 5.2 years (see Fig. 7.25).

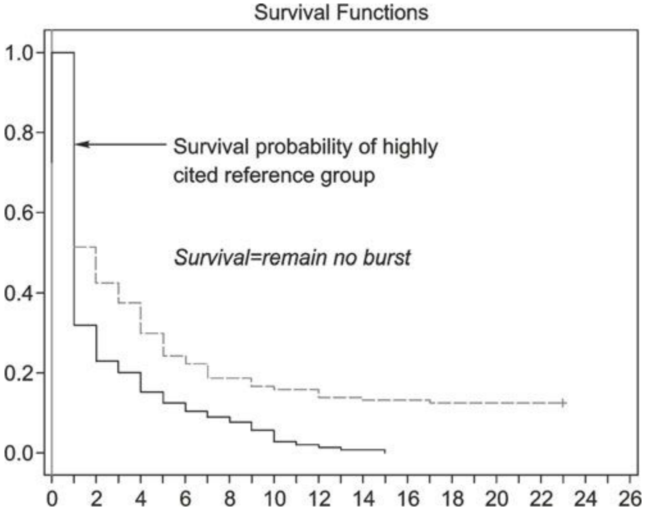


Fig. 7.25 The highly cited papers are likely to have a citation burst sooner (burst within 2.2 years) than the less cited papers (burst within 5.2 years).

7.3.3 Differentiating Awarded and Declined Proposals

The previous example suggests that survival analysis, combined with burst detection, can help us to differentiate two kinds of publications, namely papers that are highly cited and papers that are not. If we look closely at the procedure, it appears that the procedure itself is quite generic. The procedure can be used to compare not only two but several groups. A broad range of events can be defined accordingly. Thus, the procedure is applicable to compare different types of proposals, different types of patent applications, as well as different types of scientific publications. In the following example we illustrate how this procedure can be used to differentiate successful and unsuccessful proposals.

We hypothesize that awarded and declined proposals may differ in terms of how and when they deal hot topics. The timing of the appearances of a hot topic can be measured by burst detection. The comparison between the awarded and declined groups is done by survival analysis, which allows us to address the question whether the two groups differ statistically in terms of how soon hot topics appear and how long they last. The results indicate that this is indeed detectable with statistically significance with the caveat that the choice made in the noun phrase extraction and therefore the text segmentation appears to influence the sensitivity of the analysis.

We considered hypotheses that may distinguish awarded and declined proposals:

1) Awarded proposals address topics in a timelier manner than declined proposals.
2) Awarded proposals address more profound topics than declined proposals.
3) Noun phrases extracted from core segments are more specific and focused on proposed research questions than terms from one-page project summaries.
4) Survival analysis of noun phrases extracted from one-page project summaries is the same as results from noun phrases extracted from the core segments.

The first hypothesis can be tested in terms of the survival probabilities of bursts of terms as the events. Awarded proposals are expected to deal with hot topics earlier than declined ones. The second hypothesis means that the duration of a term burst in an awarded proposal is expected to be longer than the corresponding duration in a declined proposal.

The hypotheses were tested with 1,206 awarded and 4,305 declined proposals from a program of the NSF over four years (2007–2010). The results, shown in Table 7.9, include the number of noun phrases extracted and the number of noun phrases with detected bursts. The "Time till burst" column shows the *p*-level of the statistical significance between the awarded and declined groups. Notably, awarded and declined groups are distinguishable by single words and two kinds of noun phrases, namely, single-word nouns and noun phrases consisting of up to 4 words. No statistical differences were found using noun phrases with two or more words. The results suggest that awarded and declined proposals differ in their use of single words or short noun phrases.

Table 7.9 Time till bursts of terms in project descriptions of a sample of NSF proposals.

Noun phrase		Awarded			Declined			Time till burst
min	max	subtotal	bursted	(%)	subtotal	bursted	(%)	p-level
1	1	9238	1241	13.4	26358	3928	15	0.001
2	2	6018	591	9.8	21961	2159	9.8	0.990
3	3	4204	338	8%	15809	1295	8	0.760
4	4	4092	323	7.8	15492	1274	8	0.500
1	4	9541	1039	10.9	29394	3721	12.7	0.000
2	4	5743	506	8.8	21064	1939	9	0.367
single word		3963	1059	26.7	7689	3389	44	0.000

Awarded and declined proposals statistically differ in terms of the survival time till the bursts of noun phrases of 1∼4 words. Now we further compared whether the one-page project summaries provided by the principal investigators (PIs) and core passages extracted from 15-page project descriptions would be statistically different in terms of survival time till bursts of noun phrases of 1∼4 words. More detailed descriptions of how we extracted core passages will be given in Chapter 8.

We extracted core passages from 5,412 proposals (1,150 awarded and 4,262 declined). The average length of the top-1 core segments (762 words) is longer than their corresponding project summaries (613 words). The difference is statistically significant with a two-way t test, $p = 0.000$. This justified our assumption that the core information is longer than the 1-page summary but shorter than the 15-page description. In addition, one-page summaries tend to use broader terms, whereas the top-ranked core segments tend to include more specific terms. We found a statistically significant difference between the survival probabilities of awarded and declined proposals (Table 7.10).

Table 7.10. Awarded and declined proposals differ in terms of the survival time till term bursts in top-ranked core segments.

Source	Noun phrase	Awarded			Declined			Time till burst
	min/max	subtotal	bursted	(%)	subtotal	bursted	(%)	p-level
1-page summary	1/4	6815	571	8.3	23620	1990	8.4	0.916
top-1 core segment	1/4	9541	1039	10.9	29394	3721	12.7	0.000
summary/ segment		71%	55%		80%	54%		

This means that core segments are better sources of text than 1-page summaries for studies of documents such as grant proposals.

Survival analysis of bursts of noun phrases ($1 \sim 4$ words) from one-page project summaries revealed a statistically significant difference ($p = 0.007$) between awarded and declined proposals in terms of the duration of a burst. As shown in Fig. 7.26, bursts of terms in awarded proposals lasted shorter (on average 1.792 year) than in declined proposals (on average 2.381 year), although no statistically significant difference was found.

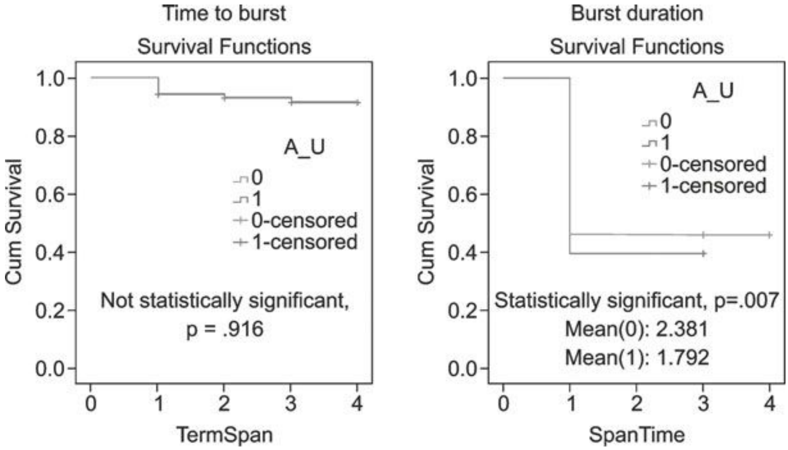


Fig. 7.26 Awarded and declined proposals have different survival probabilities of burst duration. Source: one-page project summaries of proposals (1 – awarded, 0 – declined).

7.4 Summary

In this chapter, we have first addressed some of the issues concerning how to differentiate conflicting opinions in an evidence-based approach, in particular, the role of decision trees in representing terms that may predict the orientation of customer reviews. The method particularly takes the advantage of the available ratings of reviews. Decision trees of terms and positions of reviewers provide not only a descriptive model of the central issues of a debate but also a predictive model to anticipate the paths of arguments one may follow.

The second part of the chapter deals with situations in which we do not have judgments from users or experts in numerical forms and we do not have a taxonomy or ontology either. We have introduced a method we are developing to address these issues by tapping on patterns that we can discern from linguistic relations. The flexibility needed to tolerate the ambiguity of natural language becomes available when concepts and predicates are organized in hierarchical structures.

The first part of the chapter is concerned with the role of temporal patterns such as bursts in differentiating multiple groups of documents in different categories. Survival analysis of the timing and duration of citations and term use in scientific publications and grant proposals has demonstrated the potential of this method.

References

Budiu, R., Pirolli, P., & Fleetwood, M. (2006). Navigation in degree of interest trees. http://www2.parc.com/istl/groups/uir/publications/items/UIR-2006-02-Budiu-NavigationinDOITrees.pdf. Accessed June 1, 2010.

Burt, R.S. (2004). Structural holes and good ideas. American Journal of Sociology, 110(2), 349-399.

Burt, R.S. (2005). Brokerage and closure. New York, NY: Oxford University Press.

Callon, M., Courtial, J.P., Turner, W.A., & Bauin, S. (1983). From translations to problematic networks — An introduction to co-word analysis. Social Science Information Sur Les Sciences Sociales, 22(2), 191-235.

Card, S., & Nation, D. (2002). Degree-of-interest trees: A component of an attention-reactive user interface. Proceedings of AVI (pp. 231-245).

Chalmers, M. (1992). BEAD: Explorations in information visualisation. In proceedings of the SIGIR '92(pp. 330-337). Copenhagen, Denmark. ACM Press.

Chang, C.-C., & Lin, C.-J. (2001). LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Chen, C. (2004). Searching for intellectual turning points: Progressive knowledge domain visualization. Proc. Natl. Acad. Sci. USA, 101(suppl), 5303-5310.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57(3), 359-377.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics, 3(3), 191-209.

Chen, C., Ibekwe-SanJuan, F., SanJuan, E., & Weaver, C. (2006). Visual analysis of conflicting opinions. In Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)(pp. 59-66). Baltimore, MA.

Daille, B. (2003). Conceptual structuring through term variations. In proceedings of the Proceedings of the ACL-2003 workshop on multiWord expressions: Analysis, acquisition and treatment(pp. 9-16). Saporro, Japan.

Darwin, C. (1872). The origin of species. (6th ed.): Project. Gutenberg.

Fiszman, M., Demner-Fushman, D., Kilicoglu, H., & Rindflesch, T.C. (2009). Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. Journal of Biomedical Informatics, 42, 801-813.

Fry, B. (2009). On the origin of species: The preservation of favoured traces. http://benfry.com/traces/.

Furnas, G.W. (1986). Generalized fisheye views. In proceedings of the CHI '86(pp. 16-23. ACM Press.

Ham, F.v., Wattenberg, M., & Viéas, F.B. (2009). Mapping text with phrase nets. IEEE Transactions on Visualization and Computer Graphics, 15(6), 1169-

1176.

Havre, S., Hetzler, E., Whitney, P., & Nowell, L. (2002). ThemeRiver: Visualizing thematic changes in large document collections. IEEE Transactions on Visualization and Computer Graphics, 8(1), 9-20.

Heer, J. (2007). The prefuse visualization toolkit. http://prefuse.org/.

Heer, J., & Card, S.K. (2004). DOI Trees revisited: Scalable, space-constrained visualization of hierarchical data. Proceedings of AVI (pp. 421-424).

Hetzler, B., Whitney, P., Martucci, L., & Thomas, J. (1998). Multi-faceted insight through interoperable visual information analysis paradigms. In Proceedings of the IEEE Information Visualization '98(pp. 137-144). Los Alamitos, CA: IEEE Computer Society Press.

Ibekwe-SanJuan, F. (1998). A linguistic and mathematical method for mapping thematic trends from texts. In Proceedings of the 13th European Conference on Artificial Intelligence (ECAI'98). (pp. 170-174). Brighton, UK.

Ibekwe-SanJuan, F., & SanJuan, E. (2004). Mining textual data through term variant clustering: The TermWatch system. In Proceedings of the Recherche d'Information assistée par ordinateur(RIAO 2004)(pp. 487-503). University of Avignon, France.

Kohonen, T. (1995). Self-organizing maps. Springer.

Paley, W.B. (2002). TextArc. http://www.textarc.org/.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the ACL.

PNNL. IN-SPIRE. http://in-spire.pnl.gov/.

Rip, A., & Courtial, J.P. (1984). Co-word maps of biotechnology — An example of cognitive scientometrics. Scientometrics, 6(6), 381-400.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualization. In Proceedings of the IEEE Workshop on Visual Language(pp. 336-343). Boulder, CO: IEEE Computer Society Press.

Sparck Jones, K. (1999). Automatic summarizing: Factors and directions. In I. Mani & M.T. Maybury (Eds.), Advances in Automatic Text Summarization (pp. 2-12). Cambridge, MA: MIT Press.

Swanson, D.R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, (30), 7-18.

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: Experiments with relevance and rhetorical status. Computational Linguistics, 28(4), 409-445.

Thagard, P. (1992). Conceptual revolutions. Princeton, New Jersey: Princeton University Press.

Thomas, J.J., & Cook, K.A. (Eds.). (2005). Illuminating the path: The research and development agenda for visual analytics. IEEE Computer Society Press.

Tijssen, R.J.W., & Vanraan, A.F.J. (1989). Mapping co-word structures — a comparison of multidimensional-scaling and leximappe. Scientometrics, 15(3-4), 283-295.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. Proceedings of HLT-NAACL 2003 (pp. 252-259).

Viégas, F.B., Wattenberg, M., Ham, F.v., Kriss, J., & McKeon, M. (2007). Many eyes: A site for visualization at Internet scale. IEEE Transactions on Visualization and Computer Graphics, 13(6), 1121-1128.

Wattenberg, M., & Viégas, F.B. (2008). The word tree: an interactive visual concordance. IEEE Transactions on Visualization and Computer Graphics, 14(6), 1221-1228.

Witten, I.H., & Frank, E. (1999). Data mining: Practical machine learning tools and techniques with Java implementations. San Francisco, CA: Morgan Kaufmann.

Yang, Y., & Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. In J.D.H. Fisher (Ed.), Proceedings of the The 14th International Conference on Machine Learning (ICML’97)(pp. 412-420). Nashville, US. Morgan Kaufmann Publishers.

Chapter 8 Transformative Potential

Identifying and supporting high-risk and high pay-off research has been one of the major concerns of science policy as well as individual scientists and their institutions. The National Science Foundation (NSF) has been concerned about identifying and funding transformative research for decades. The U.S. is not the only country that is experiencing the sense of urgency like the Gathering Storm we discussed at the beginning of the book. Research Councils UK (RCUK) considers high potential, high impact research as adventurous, speculative, innovative, exciting, creative, radical, groundbreaking, precedence setting, unconventional, visionary, challenging, ambitious, uncertain, mould-breaking or revolutionary (RCUK 2006). The Natural Sciences and Engineering Research Council of Canada (NSERC) defined the concept of risk based on unconventionality and an uncertainty of results (NSERC 2003).

Due to the intensified international competition, shrinking public funds, and increasingly stringent criteria imposed by funding agencies, obtaining competitive research funding is now a common problem worldwide at various levels. Scientists have to make tough decisions (NSF, 2007) and balance the precious time and effort on writing research grants that have a diminishing chance of success and face increasingly overwhelmed reviewers. While many funding agencies encourage high-risk and high payoff research, assessing the transformative potential of a research idea is an increasingly intensified challenge.

8.1 Transformative Research

Transformative research is scientific work that revolutionizes a topic, a field of study, and even a discipline. There are several common types of creative accomplishment such as new theories, new discoveries, new methodology, new instruments, and new syntheses.[1] Scientific revolutions are transformative research. Scientific breakthroughs are transformative research. High-risk and

[1] http://www.cherry.gatech.edu/crea/

C. Chen, *Turning Points*

high pay-off ideas are potentially transformative research. It becomes less clear to distinguish the research that is positioned between the high-end of transformative research and 'traditional' research. There is a consensus following TRACES and more recent assessments that scientific breakthroughs could only be identified after extended periods. The National Science Board pointed out that the main problem in financing new transformative projects is the lack of faith on the part of applicants.

Scholars in Europe refer to transformative research as breakthrough research and mould-breaking research (Härynen, 2007). Transformative research is characterized by its potential of tackling exceptionally wide and complex research problems, challenging established theories and scientific paradigms, and introducing radically new ways of using methods and unprejudiced combination and interdisciplinary integration of different research perspectives. Breakthrough research is characterized by an exceptional risk of failure. Indeed, it is commonly believed that an overly conservative peer reviewing system is one of the key obstacles to achieving these breakthroughs.

Is there a point beyond which research suddenly becomes non-transformative? How soon can we expect to recognize transformative research that has accomplished years ago or a newly proposed research that has transformative potential?

The Academy of Finland's annual research funding amounts to more than 260 million Euros, which is around 15% of the Finnish government's annual R&D spending. In early 2006, the Academy of Finland commissioned an inquiry into the nature of breakthrough research, the need for funding it and its funding criteria. In particular, the Academy was seeking answers to two major questions:

1) how breakthrough research can be identified in the project proposal review process;
2) how research funding should encourage the propagation of new ideas and the taking of risks.

The Academy commissioned the inquiry to Maunu Häyrynen, who started the work by discussing with senior management at the Academy, the heads of Research Units at the Academy's Administration Office and other key personnel. Häyrynen conducted a detailed analysis of proposals submitted to the Academy of Finland for general research grants in 2005 in selected fields of research under each of the four Research Councils. The report was published in 2007 (Härynen, 2007).

The Häyrynen report concluded that there is as yet no agreement on how to identify breakthrough research and how it should be encouraged by means of research funding, but identified some options:

- Allocate dedicated funding to research issues of current interest or strategic priority areas;
- Apply a separate set of criteria for breakthrough research;
- Give adequate recognitions of breakthrough research, both successful and failures;

- Modify criteria to favor innovation and tolerate risks.

The Häyrynen report described a 2004 audit of the Academy of Finland. The audit was primarily based on interviews with the Academy's President, Vice Presidents (Research and Administration), and Board members. The interviews identified possible hallmarks of transformative research as follows:

1) Research that consists of exceptionally innovative basic research;
2) A well-established researcher moving to a new field of inquiry where they have no track records;
3) The application of a set of methods from one discipline to a completely different field;
4) The development of new methods and technologies that may not have near-term applications;
5) Research that may overthrow prevailing theory;
6) New and untested ideas;
7) Research that requires the collaboration of different disciplines.

The track record of the Academy of Finland in handling the perceived risks of proposals was mixed. On the one hand, the Academy had rejected projects that later on became successful stories with funding from elsewhere. On the other hand, the Academy had funded projects of considerable risks and projects that had never produced results of any scientific value. In terms of the role of researchers' age, it has been argued that both experienced researchers moving to a new field and young researchers should be awarded regardless their age, although less stringent criteria should be used to assess younger researchers than senior researchers.

The Häyrynen report described a survey of a total of 206 proposals received by the four Research Councils of the Academy of Finland for general research grants in 2005. The four research councils are for biosciences and environment, health, culture and society, and natural sciences and engineering. According to the report, "Based on the first screening it was clear that it could not be inferred from proposals that received weak reviews (a grade of $1 \sim 2$ on a scale from 1 to 5) or from the expert opinions on those proposals whether or not the proposed project represented breakthrough research." In other words, one cannot identify breakthrough research from either the proposals or their reviews in a pool of proposals of varying quality. The final analysis narrowed down to proposals with ratings of $3 \sim 5$ only and focused on conflicting evaluations between preliminary reviews and the final panel evaluation. Terms frequently used by reviewers to describe exceptionally innovative projects include original, novel, unique, forefront, innovative, exciting, transformative, cutting edge, and ambitious. Reviewers' statements concerning risks were categorized into seven types, namely, risks related to research objectives, to research methods, to the field of research, to personnel, to disciplinarity, to resources, and to ethics.

In terms of the funding decisions, proposals rated as exceptionally innovative and ambitious were likely to be funded, whereas those rated as involving high risks were likely to be rejected. The Häyrynen report concluded that breakthrough research is about searching for alternative paths of development and it should serve as a necessary counterbalance to science policy steering that is based on foresight and research indicators.

8.2 Detecting the Transformative Potential

We develop a number of generic metrics that can be used to identify the transformative potential of a given research idea, in particular, in connection with network representations of contemporary knowledge. Assume at a given time point t, the knowledge of a topic, a field, or a discipline up to that point $K(t)$ can be represented by mixtures of various subtopics or associative networks of conceptual components and their interrelationships. For any scientific ideas either as topics or conceptual components emerged after the time point t, i.e. $t + \Delta t$, measure the extent to which these new ideas depart from $K(t)$, the accumulated knowledge up to t.

In this chapter, we demonstrate how this type of metrics can identify potential transformative ideas in terms of the extent to which these ideas contribute to structural and/or topical variations of $K(t+\Delta t)$ with reference to $K(t)$. In other words, we conceptualize that we will be able to identify the nature of transformative research in terms of structural and topical variations over time if we can measure the distance or divergence of an updated representation of knowledge from a baseline representation. In the rest of this chapter, we mainly focus on network representations as one of the many potentially valuable routes to achieve our goals. For example, the degree of divergence of collective knowledge over time Δt can be similarly derived and measured based on topical models of a body of relevant scientific publications.

There are a number of reasons why we choose to focus on structural variations in network representations. The first reason is theoretical. Our explanatory theory of discovery suggests that one of the mechanisms for advancing scientific knowledge is to make novel connections between previously disjoint bodies of knowledge. This theory means that we can measure the novelty of newly proposed connections in terms of the extent the new links are positioned between previously disjoint bodies of knowledge. If a new link unprecedentedly connects two or more distant fields of study, then its novelty measure should be high. In contrast, if a newly added link merely repeats an existing link, then its novelty measure should be low. Between the two extremes, a newly proposed link may introduce new interpretations of existing evidence without introducing structural variations as far as the specific knowledge representation is concerned. Second, from the perspective of foraging for knowledge, the perceived profitability is in line with the high-risk and high pay-off expectation of transformative research; we would give higher scores to those novel and unprecedented connections because those conceptualizations are harder to come by. Third, detecting structural variations in networks allows us to pin point the specific links that alter the structure of existing knowledge the most, which is valuable information for additional validation, for example, by consulting with scientists themselves and other domain experts. In addition, the ability to pinpoint the potential of specific connections makes it possible for analysts to keep track the evolution of their impact over time so that one can verify whether scientific ideas that are identified today with transformative potential are evidently transformative as shown in due course.

In order to assess the extent that these metrics can capture transformative research, our strategy is to take a retrospective-predictive approach by predicting citation counts received by scientific publications that had induced strong structural variations in the past. In other words, our hypothesis is that the degree of structural variation introduced by a scientific publication (as a symbol of scientific ideas) is a significant predictor of its citation counts in subsequent years.

8.2.1 Connections between References and Citations

Fig. 8.1 shows two plots related to original research articles on the subject of mass extinction between 1975 and 2010. What is interesting is that the two plots appear to run in parallel to each other most of the time until the most recent years. Is it something purely coincident? Does it also happen to other subjects?

Before we address these questions, we need to introduce some notations and terminologies. We use the notation $d_{\text{source}} \rightarrow d_{\text{target}}$ to denote the fact that a scientific publication d_{source} cites another scientific publication d_{target}. For a given scientific publication d, its references R are scientific publications it cites, i.e. references $(d) = \{r_i | d \rightarrow r_i\}$, whereas citations are the instances that subsequent scientific publications cite d, i.e.. citations $(d) = \{c_i | c_i \rightarrow d\}$. The two plots in Fig. 8.1 trigger a hypothesis that the number of references made by an article is correlated with the number of citations it receives. In other words, articles with a longer list of references appear to receive more citations than articles with a shorter list of references.

A news article[2] on Nature News published on August 13, 2010 was quick to jump to the conclusion that an easy way to boost a paper's citation is to include more references. Gregory Webster, the psychologist at the University of Florida in Gainesville, found a strong correlation between the number of references and the number of citations based on 50,000 Science papers but he made a superficial claim, "if you want to get more cited, the answer could be to cite more people." First, the claim stretched a simple correlation to a causal relationship. Second, the claim lacked the support of a theory that

[2] http://www.nature.com/news/2010/100813/full/news.2010.406.html

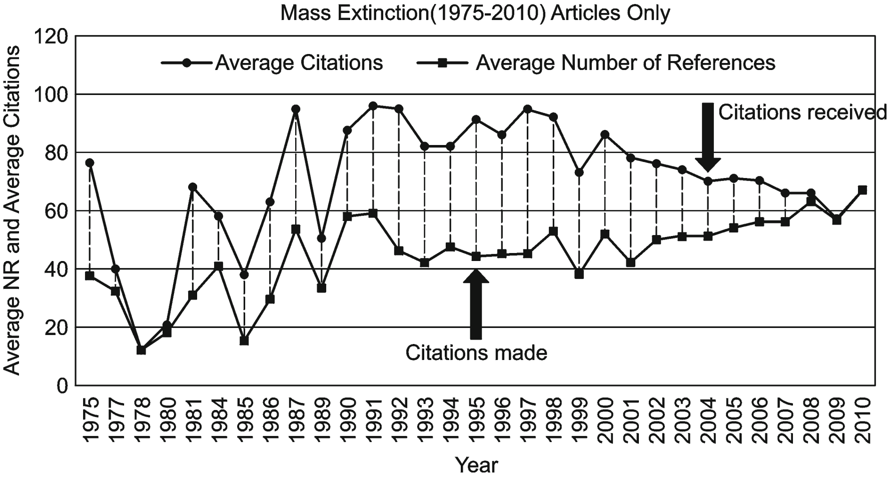


Fig. 8.1 A correlation between the average number of references of articles and their average citations.

can explain why this would be the case. A few informetricians questioned the claim. One of them, Ronald Rousseau, the President of the International Society for Scientometrics and Informetrics (ISSI), put forward a conjecture: an article that deals with several topics has a higher probability of being useful and being cited than an article that is relevant to just one subfield. The number of references itself is not the cause of the relationship. Rousseau asked if anyone can prove or disprove the conjecture.

In fact, Rousseau's conjecture is expressed almost in a form that can be derived from our explanatory theory of transformative discovery. According to our theory, the brokerage mechanism is one of the key mechanisms that can lead to transformative research. The brokerage mechanism, also known as boundary spanning, creates unprecedented links that connect previously disjoint topics or bodies of knowledge. A paper on potentially transformative discovery is then likely to build conceptual bridges between multiple topics and even distinct fields or disciplines. By doing so, it becomes natural that it tends to cite references from multiple topics or fields and, as a result, leads to a higher total number of references than a paper on a single topic would cite. More significantly, due to the transformative value of the paper, it is likely to receive more citations than less transformative papers. Therefore, we hypothesize that it is not the total number of references that causes a higher citation count; rather, high citations are much more likely to be caused by the structural variation introduced by the transformative paper.

The next step is to test this hypothesis. The most straightforward strategy is to first compute structural variation metrics of scientific papers published after time t with reference to $K(t)$, the knowledge structure up to time t, then test to what extent such metrics predict subsequent citations obtained by these papers for the next 5 or 10 years alongside with other variables that have been commonly considered as predictors of citation counts, such as the length of a paper, the number of collaborating authors of a paper, and the number of references of a paper.

In fact, many of these variables are derivable from the central hypothesis that unusual conceptual linkages are likely to indicate the transformative potential. For example, the quality of a paper may be not really related to the number of its co-authors; instead, it may be related to the number of topical areas where coauthors come from. Similarity, the quality of a paper may be not related to the number of countries of coauthors, but, instead, related to the number of distinct disciplines the coauthors belong to.

One of the essential criteria of a good theory is its coherence. That is whether the theory can explain many seemingly different things under the same framework. In the case of our explanatory theory of transformative discovery, the theory has largely reduced the number of variables to consider with the same underlying explanation. We introduce the rationale of the design of metrics of structural variation in the following sections.

8.2.2 Measuring Novelty by Structural Variation

The structure of a network characterizes the connectivity and interrelationship of a set of entities. An important question concerning the dynamics of a network is whether the stability of its structure will be affected as new information becomes available. In practice, every network has its own context, or environment. If changes take place in its environment, it is often necessary to ask whether these changes in the environment have any impact on the structure of the network. In light of new information or new evidence, it may become necessary to update the structure of the existing network, especially when it may lead to a significantly different structure.

The concept of structural variation can be illustrated in terms of the strength of an expected association between two concepts. Consider the following examples of word associations. Which one do you take for granted and which one would surprise you?

1) soccer $\sim$ beer
2) soccer $\sim$ octopus

The connection between soccer and beer seems to be widely known — every four years football fans all over the world watch FIFA's world cup matches and many of them would go and watch the games in bars, pubs, or anywhere near to beers. In contrast, the word *octopus* had little to do with the sport of soccer — until 2010 FIFA world up in South Africa; an octopus was under the spotlight before more and more important games because it mysteriously managed to predict the winners of so many games so accurately. The connection between soccer and octopus was novel because prior to the 2010 World Cup, most people would not think of an octopus in any association with soccer.

Similar examples are also available in science and technology. Before terrorist attacks on September 11, 2001, the literature of post-traumatic stress disorder (PTSD) generally focused on people who were on site when traumatic events took place. After 911, however, researchers realized that even people who were not anywhere physically near to a trauma could still develop symptoms of PTSD due to graphical news coverage and extensive special coverage on mass media:

1) people $\sim$ eyewitness/experience trauma $\sim$ PTSD
2) people $\sim$ news coverage on mass media $\sim$ PTSD

A search for potentially transformative research can be at least partially fulfilled by finding scientific papers that make such structural changes to the intellectual structure of a subject domain, e.g. PTSD in this example. Then our hypothesis becomes that papers making this type of contributions are more likely to be cited than papers making less significant structural changes.

In the history of science, there are many examples of how new theories revolutionized the contemporary knowledge structure. For example, the 2005 Nobel Prize in medicine was awarded to the discovery of Helicobacter pylori, a bacterium which was not believed to be possible to find in human's gastric system (Chen, Chen, Horowitz, Hou, Liu, & Pellegrino, 2009). In literature-based discovery, Swanson discovered previously unnoticed linkage between fish oil and Reynaud's syndrome(Swanson, 1986). In drug discovery, one of the major challenges is to find new compound structures effectively in the vast chemical space that satisfy an array of constraints(Lipinski & Hopkins, 2004). In mapping scientific frontiers(Chen, 2003) and studies in science of science (Price, 1965), it would be particularly valuable if scientists, funding agencies, and policy makers can have tools that may assist them to assess the novelty of ideas in terms of their conceptual distance from the contemporary domain knowledge. In these and many more scenarios, a common challenge for coping with a constantly changing environment is to estimate the extent to which the structure of a network should be updated in response to newly available information.

The metrics to be introduced below are generic and suitable for a variety of networks. To illustrate the use of such metrics, we focus on intellectual networks of scientific domains and show how these metrics can be used to detect potentially significant new publications.

A document co-citation network $\boldsymbol{G}(V, E)$ can be generated from a set of scientific publications S. Each node n, a member of V, in such a network represents a scientific publication cited by a member of the given set S. An edge e_{ij} connecting nodes n_i and n_j in the network represents a co-citation relationship, which means if there exists s in S such that s cites both n_i and n_j. Usually an edge is weighted to reflect the relative strength of such binding. The more often such co-citation instances there are, the stronger the edge grows between them. We are interested in the following question: given a new publication s' arrived from a new set of publications S', what structural changes does s' introduce with regard to the $\boldsymbol{G}$ formed prior to the arrival of s'. In other words, we seek to measure δE, the change of E to E' in the new $\boldsymbol{G}(V, E')$. Note that V remains constant. Here, we limit our focus to situations of constant Vs. Scenarios in which V varies are more complex; they will need to be addressed once we have a better understanding of the relatively simple cases of a constant V.

In essence, we define structural change metrics in terms of δE with respect to E. The simplest metric is $|\Delta E|$, the number of different edges introduced by the new s'. If the new paper s' uniformly cites all the references in the existing network G, then s' is adding nothing new to the structure of the network, thus $|\Delta E| = 0$. If all the references cited by the new paper s' already exist in E, then it is adding very little to the structure of the network as far as the network topology is concerned, although it in effect reinforces a substructure of the network. If s' adds new edges to the original network, then we are receiving new information that may potentially lead to a global change of the network.

A more sophisticated metric takes the position of each node in the network into account. For example, a metric can be defined according to the change of centrality scores of all the nodes in the network. The node centrality of a network $\boldsymbol{G}(V, E), C(\boldsymbol{G})$ is a distribution of the centrality scores of all the nodes, $< c_1, c_2, \ldots, c_n >$, where c_i is the centrality of node n_i, and n is $|V|$, the total number of nodes. The degree of structural change δE can be defined in terms of the K-L divergence, we denote this metric as $\Delta_{\text{centrality}}$.

The next metric $\Delta_{\text{modularity}}$is defined to measure the novel associations added across aggregations of nodes. First, decompose $\boldsymbol{G}(V, E)$ to a set of clusters, $\{C_k\}$; in this case, C_k is a co-citation cluster (Chen, Ibekwe-SanJuan, & Hou, 2010). Given a cluster configuration, the modularity of the network can be computed. The modularity measures whether the network can be decomposed nicely with the given clusters. A high modularity means that the given cluster configuration can divide the network into relatively independent partitions with few cross cluster edges. In contrast, a low modularity means that the given cluster configuration cannot divide the network without many cross-cluster edges. If a new paper s' adds an edge connecting members of the same cluster, it will have no impact on the modularity. It will not make any difference to the value of $\Delta_{\text{modularity}}$. On the other hand, if s' adds an edge between different clusters and the two clusters are previously not connected, the modularity of the new structure will be lower than that of the original structure. $\Delta_{\text{modularity}}$=modularity($\boldsymbol{G'}$)/modularity($\boldsymbol{G}$).

The modularity of a network is a function of a set of alternative partitions of the network. Some partitions lead to a higher modularity, whereas others lead to lower modularity scores. The optimal partition can be determined based on the variation of modularity scores over different partitions of the same network. Since the maximum modularity implies the maximum separation of various network components, it is often used as a criterion to choose the corresponding clusters as the most representative solution. For example, if a co-citation network with an 8-cluster solution has a modularity of 0.4838, its modularity is lower either with more clusters (0.4791 for 9 clusters) or less clusters (0.3355 for 7 clusters), then it is reasonable to choose the 8-cluster solution as the basis for the calculation of $\Delta_{\text{modularity}}$.

Fig. 8.2 illustrates how structural changes made by a newly arrived article #3 to an existing network on fish oil and Reynaud's syndrome. Article #3 added a new connection between articles #1 and #2 by citing both of them together. The new connection led to a reduction of the modularity by 0.022% and a normalized change of centrality by 0.016%. Article #3 was late cited 14 times. Our hypothesis is that statistically, the modularity and centrality change metrics could account for the number of citations as an estimate of its significance.

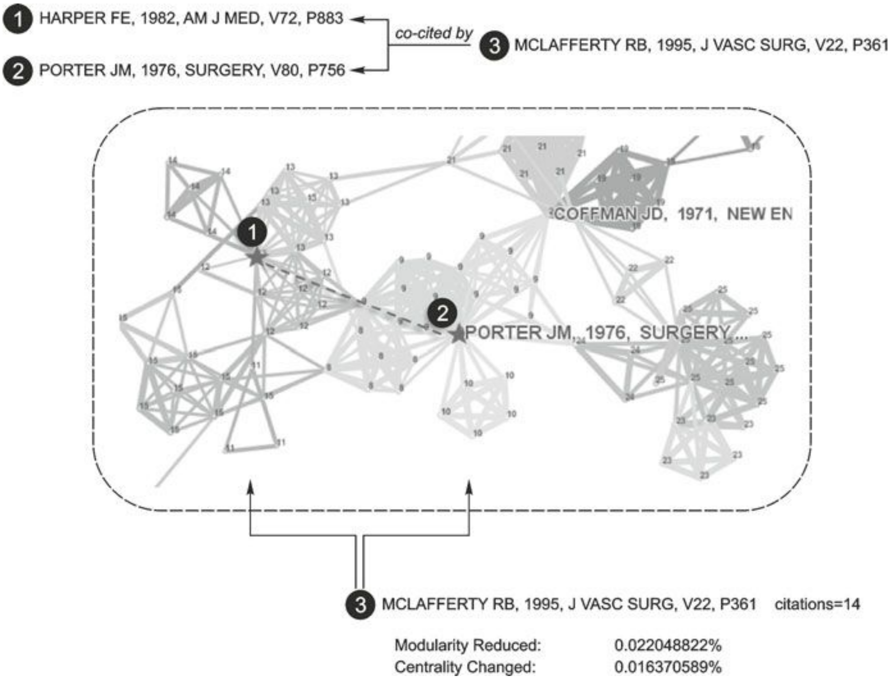

1 HARPER FE, 1982, AM J MED, V72, P883
co-cited by
3 MCLAFFERTY RB, 1995, J VASC SURG, V22, P361
2 PORTER JM, 1976, SURGERY, V80, P756
COFFMAN JD, 1971, NEW EN
1
2
PORTER JM, 1976, SURGERY...
3 MCLAFFERTY RB, 1995, J VASC SURG, V22, P361 citations=14
Modularity Reduced: 0.022048822%
Centrality Changed: 0.016370589%

Fig. 8.2 An incoming article #3 added a new link that connects references #1 and #2. The accumulated change of modularity by #3 is 0.022% with respect to the network structure without connections made by #3. Data source: Fish oil and Reynaud's Syndrome.

Fig. 8.3 shows that the same method is applicable to detect the novelty of a paper with reference to a network of co-occurring terms found in previous publications on terrorism research. A 2001 paper by P. J. Maddox made a fresh connection between terms *terrorist attacks* and *accidental exposure*, which led to 0.108% of reduction in modularity and 0.002% of divergence in centrality. An even stronger change of structure was made by Y. Matsuda's 1998 paper because it linked terms *terrorist attack* and *Tokyo subway*, suggesting that *Tokyo subway* is probably a topic that had not appeared in the literature in the context of terrorism.

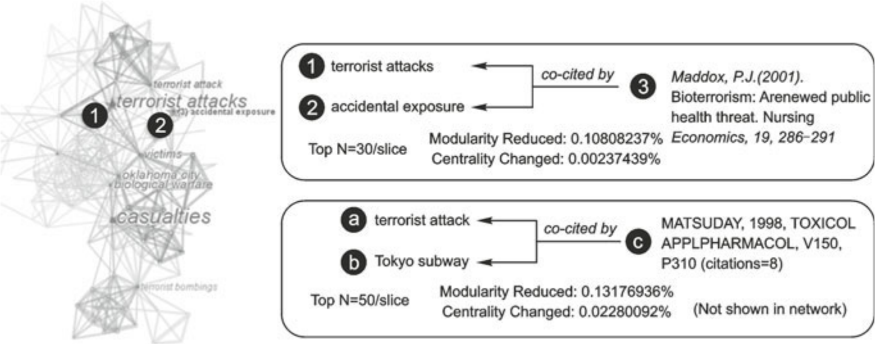


Fig. 8.3 The novelty of a newly arrived paper can be also determined according to the changes in modularity and centrality in a network of terms. Data source: Terrorism Research.

8.2.3 Statistical Validation

In order to verify what these metrics measure, we hypothesize that articles would be ranked high by these metrics if they contribute to novel connections to the existing structure of a network. Thus we may be able to use these metrics to construct a way to detect the novelty of new papers. According to the explanatory and computational theory of discovery we have developed, potentially significant scientific discoveries tend to be boundary spanning work across different patches of previous scientific knowledge (Chen et al., 2009). The metrics defined earlier in this chapter are likely to reflect the novelty of ideas in the newly arrived papers because the higher the values of the structural change metrics are, the more likely the new papers are making novel connections that are not captured by the current structure of the network. Furthermore, if a new paper is making novel contributions, it is reasonable to hypothesize that it will be in a good position to attract more citations later on than papers making less novel contributions.

To test this hypothesis, we need to verify that the structural change metrics of new papers are good predictors of how many citations they will receive later on. First, we test the hypothesis using a simple ANOVA and then using a negative binomial regression model.

For the UNIANOVA test, structural variation metrics $\Delta_{\text{modularity}}$ and $\Delta_{\text{centrality}}$ are used as the co-variant variables to predict the dependant variable *Citations*. We also included two additional scores *alpha* and *beta*, where *alpha* is the proportion of existing and redundant edges made by the article in question and beta is the proportion of new edges added by the article. We controlled the effect of *NR*, the total number of references of the article. The dependent variable *Citations* is the number of citations the article has been cited up to 2010.

```
UNIANOVA Citations WITH Δmodularity   Δcentrality alpha beta
  /REGWGT=NR
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PRINT=PARAMETER ETASQ
  /CRITERIA=ALPHA(.05)
  /DESIGN=Δmodularity   Δcentrality alpha beta.
```

The hypothesis was tested on papers that cite a major paper on CiteSpace (Chen, 2006) in the Web of Science because we have extensive expertise in relevant areas of research so that we could draw upon our domain knowledge in the analysis and interpretation of the results. The data source contained 76 articles, including 32 journal papers, 38 conference proceeding papers, 5 review papers, and 1 editorial. They were written by a total of 229 authors from 108 institutions. This dataset represents a total of 3,647 references. The distributions of these articles from 2006 through 2010 are: 1, 17, 16, 31, and 11. The primary subject category of these articles is information science and library science. The secondary one is computer science, information systems, and interdisciplinary applications.

The algorithm operates as follows. At any time point t in the interval [2006, 2010], the goal is to estimate the novelty of papers published in year t, S_{t}, according to how much structural changes they introduced in comparison with the network structure up to the year before. In other words, the algorithm constructs a network of co-cited references $\boldsymbol{G}_{t-1}$ based on papers published in the interval [2006, $t-1$]. $\Delta_{\text{modularity}}(s)$ and $\Delta_{\text{centrality}}(s)$ are computed for each $s \in S_t$. For example, for t=2009, there are 31 papers in the incoming stream. Their novelty is computed with reference to the co-citation network formed based on references cited by 1+17+16=34 papers published prior to 2009. In CiteSpace, we generated networks per slice with the top 200 most cited references in each time slice. Further investigations are needed to find out the impact of such selection criteria on the final ranking results. Table 8.1 lists the details of accumulative networks prior to a given

Table 8.1 The accumulative networks prior to the streaming articles.

1-year slices	citers	criteria	space	nodes	links	networks	size	modularity
2006	1	top 200	19	19	171	G_{2006}	19×19	0.0000
2007	17	top 200	338	200	2634	G_{2007}	216×216	0.7340
2008	16	top 200	1526	200	9261	G_{2008}	399×399	0.2268
2009	31	top 200	868	200	2432	G_{2009}	558×558	0.3269
2010	11	top 200	475	200	2933			

time point t.

Takeda and Kajikawa (2010) studied the change of modularity in networks of direct citations and found that the evolution of such direct citation networks appears to have three stages. Core clusters are first formed, followed by peripheral clusters, and then by the further growth of the core clusters. Taleda and Kajikawa adopted the clustering algorithm originally introduced by Mark Newman. Newman's algorithm starts with a bottom-up procedure in which individual nodes are joined together. The algorithm searches for the structure that represents the maximum modularity. Instead of searching for the maximum modularity, Taleda and Kajikawa simply kept track of the modularity at each step of the process and used this information to explore the structural change in the network. In our analysis, the network in 2006 was formed by the citation behavior of one article. By 2007, the network represented the citation trails of 18 articles with a very high network modularity of 0.7340, suggesting a relatively clear partition of the network into distinct topics. By 2008, the network grew even larger with contributions from additional 16 articles. The modularity of the new network dropped to 0.2268, which means the overall interconnectivity of the network was increased considerably. Finally, by incorporating co-cited references from another 31 articles, the network now contained 558 references. Interestingly, as the network continued to grow, the modularity increased to 0.3269, suggesting that new topics were probably introduced into the network and the boundaries between the new topics and older ones are still recognizable.

Now let's look at the ranking results in Table 8.2. If a paper scores based on whether it adds novel connections between clusters in a network, it follows that the paper creates bridges or boundary-spanning links between previously unconnected patches of knowledge. What types of papers would be ranked high in such scenarios? Fig. 8.5 shows a ranked list of papers that cited (Chen, 2006) by $\Delta_{\text{modularity}}$, which is the first column in the table.

Table 8.2 Top-10 papers ranked by the modularity variation rate ΔQ, i.e. $\Delta_{\text{modularity}}$

ΔQ	ΔC	TC	NR	Author	Year	Title	Source
4.5329	.0567	18	610	JUDIT BARILAN	2008	Informetrics at the beginning of the 21st century — A review	J INFORMETR
2.0735	.0236	3	370	STEVEN A. MORRIS	2008	Mapping research specialties	ANNU REV INFORM SCI TECH
1.5902	.0044	3	106	CHAOMEI CHEN	2009	Towards an explanatory and computational theory of scientific discovery	J INFORMETR
.8241	.0024	1	62	ERJIA YAN	2009	Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis	J AM SOC INF SCI TECHNOL

Continued

ΔQ	ΔC	TC	NR	Author	Year	Title	Source
.7701	.0014	2	29	YOSHIYUK I TAKEDA	2009	Optics: a bibliometric approach to detect emerging research domains and intellectual bases	SCIENTOM ETRICS
.7079	.0037	1	84	KATY BORNER	2009	Visual conceptualizations and models of science	J INFORM-ETR
.4769	.0003	0	23	YOSHIYUK I TAKEDA	2010	Tracking modularity in citation networks	SCIENTOM ETRICS
.4635	.0026	1	45	YOSHIYUK I TAKEDA	2009	Nanobiotechnology as an emerging research domain from nanotechnology: A bibliometric approach	SCIENTOM ETRICS
.4124	.0008	0	42	ALEKS ARIS	2009	Visual Overviews for Discovering Key Papers and Influences Across Research Fronts	J AM SOC INF SCI TECHNOL
.3574	.0012	0	33	ERJIA YAN	2009	The Use of Centrality Measures in Scientific Evaluation: A Coauthorship Network Analysis	PROC INTER CONF SCI INFOMET

Based on our expertise in the domain, we immediately recognize that the first and second articles are in fact review articles. Review articles satisfy the expected patterns because they tend to survey and review a wide range of topics, which would explain why they stand out when we focus on the behavior of adding new co-citation links across different topics. The third position is our own paper. It is not a review paper; however, it covers a number of distinct topics in a new framework of an explanatory and computational theory of discovery. It cited 106 references.

Recall there are 5 review articles in the dataset. Where are the other three review articles? Mike Thelwall's review of bibliometrics and webometrics is ranked 21st in the list. A conjecture is that his review focused on the two reasonably well connected topics in a more diverse overall context. Another interesting example is the article by Katherine McCain in 2008. Her article cited 282 references. It is possible that, due to the specific topic of her article — the oeuvre of Conrad Hal Waddington — many of her references may have little overlap with the references cited by other papers in the dataset.

In order to better understand why non-review articles are highly ranked, we examined the exact co-citations that were added to the growing network for the first time by non-review articles. For example, one of the 'unusual' connections made by our JOI paper (the third paper written by Chaomei

Chen in Table 8.2) is the connection between Diana Crane's work on invisible college along with an article by K. Dunbar on scientific discovery. Similarly, the co-citation added by the 4th paper written by Erjia Yan in Table 8.2 between Freeman's paper on centrality and the h-index paper by Hirsch is among the 'unusual' ones as far as this dataset is concerned. In another example, the article by Yoshiyuki Takeda in 2009, the 5th paper in Table 8.2, co-cited Klavans_2006 and Chen_2002. In summary, our new metrics provide a holistic measure of the 'unusual' connections contributed by a new paper.

We tested our hypothesis with a univariate General Linear Model. The results are shown in Tables 8.3 and 8.4. The model found a statistically significant effect of $\Delta_{\text{Centrality}}$ in predicting the number of citations ($p = 0.007$), but no significant effect was found for $\Delta_{\text{Modularity}}$. The model explained 87.5% of the variance, thus it can be regarded as a sufficiently accurate model. Table 8.4 indicates that the effect of the centrality divergence is practically meaningful.

Table 8.3 Tests of Between-Subjects Effects[b]. Data source: 76 papers citing (Chen, 2006).
Dependent Variable: Citations

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	112675.351[a]	4	28168.838	58.578	.000	.890
Intercept	2331.753	1	2331.753	4.849	.036	.143
ΔModularity	801.177	1	801.177	1.666	.207	.054
ΔCentrality	**4098.399**	**1**	**4098.399**	**8.523**	**.007**	.227
alpha	46.711	1	46.711	.097	.758	.003
beta	1263.181	1	1263.181	2.627	.116	.083
Error	13945.494	29	480.879			
Total	214646.000	34				
Corrected Total	126620.845	33				

a. R Squared = .890 (Adjusted R Squared = .875)
b. Weighted Least Squares Regression – Weighted by NR

Table 8.4 Parameter Estimates[a]
Dependent Variable: Citations

Parameter	B	Std.Error	t	Sig.	95% Confidence Interval		Partial Eta Squared
					Lower Bound	Upper Bound	
Intercept	1.541	.700	2.202	.036	.110	2.971	.143
ΔModularity	4.861	3.766	1.291	.207	–2.841	12.564	.054
ΔCentrality	**594.105**	**203.504**	**2.919**	**.007**	**177.891**	**1010.318**	**.227**
alpha	.011	.035	.312	.758	–.061	.083	.003
beta	–.210	.130	–1.621	.116	–.476	.055	.083

a. Weighted Least Squares Regression – Weighted by NR

The results are encouraging and we expect that these metrics can provide valuable information needed in the analysis of the dynamics of networks and dealing with changes and uncertainties. Next, we tested the hypothesis with a negative binomial regression.

Negative binomial regression models are frequently used in the literature when analyzing frequency data that the mean is much smaller than the variance. Paper citations and patent citations are a typical type of count data. Various studies have used ordinary linear regression models. However, researchers also notice that citation data tend to have many zeros and small values. In other words, the variance in citation data is often greater than the mean. Negative binomial regression models are more appropriate to model this type of data (Lee, Lee, Song, & Lee, 2007; Lokker & Walter, 2010). The negative binomial regression was specified as a generalized linear model with modularity and centrality variation rates as co-variant variables to predict the number of citations retrospectively.

```
* Generalized Linear Models.
GENLIN Citations WITH ModularityVariation CentralityVariation
  /MODEL ModularityVariation CentralityVariation INTERCEPT=YES
OFFSET=Year SCALEWEIGHT=NR
 DISTRIBUTION=NEGBIN(1) LINK=LOG
   /CRITERIA METHOD=FISHER(1) SCALE=1 COVB=ROBUST
MAXITERATIONS=100 MAXSTEPHALVING=5 PCONVERGE=1E-006(ABSOLUTE)
SINGUL-AR=1E-012 ANALYSISTYPE=3(LR) CILEVEL=95 CITYPE=WALD
LIKELIHOOD=FULL
   /MISSING CLASSMISSING=EXCLUDE
   /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION.
```

The results of statistical tests are shown in Tables 8.5∼8.7. The model effects of both modularity and centrality variation rates are statistically significant in predicting citation counts. In terms of parameter estimates, the centrality variation rate is statistically significant, but the modularity variation rate is not. The result of the negative binomial regression is consistent with the results of the UNIANOVA test.

Table 8.5 Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
3892.663	2	.000

Dependent Variable: Citations
Model: (Intercept), Modularity Variation, Centrality Variation, offset = Year
a. Compares the fitted model against the intercept-only model.

Table 8.6 Tests of Model Effects

Source	Type III		
	Likelihood Ratio Chi-Square	df	Sig.
(Intercept)	.[a]		.
ΔModularity Variation	128.181	1	.000
ΔCentrality Variation	303.783		.000

Dependent Variable: Citations
Model: (Intercept), ModularityVariation, CentralityVariation, offset = Year
a. Unable to compute due to numerical problems

Table 8.7 Parameter Estimates

Parameter	B	Std.Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	–2007.985	.5873	–2009.136	–2006.834	1.169E7	1	.000
ΔModularity Variation	–.715	.4460	–1.589	.159	2.569	1	.109
ΔCentrality Variation	**73.305**	**29.3452**	**15.789**	**130.821**	**6.240**	**1**	**.012**
(Scale)	1^{a}						
(Negative binomial)	1						

Dependent Variable: Citations
Model: (Intercept), ModularityVariation, CentralityVariation, offset = Year
a. Fixed at the displayed value.

Fig. 8.4 is an example of what types of papers would be ranked high by the modularity variation metric and the centrality variation metric. The papers cited a 2006 journal paper on CiteSpace II (Chen, 2006). The x-axis is the modularity variation rate and the y-axis is the centrality variation

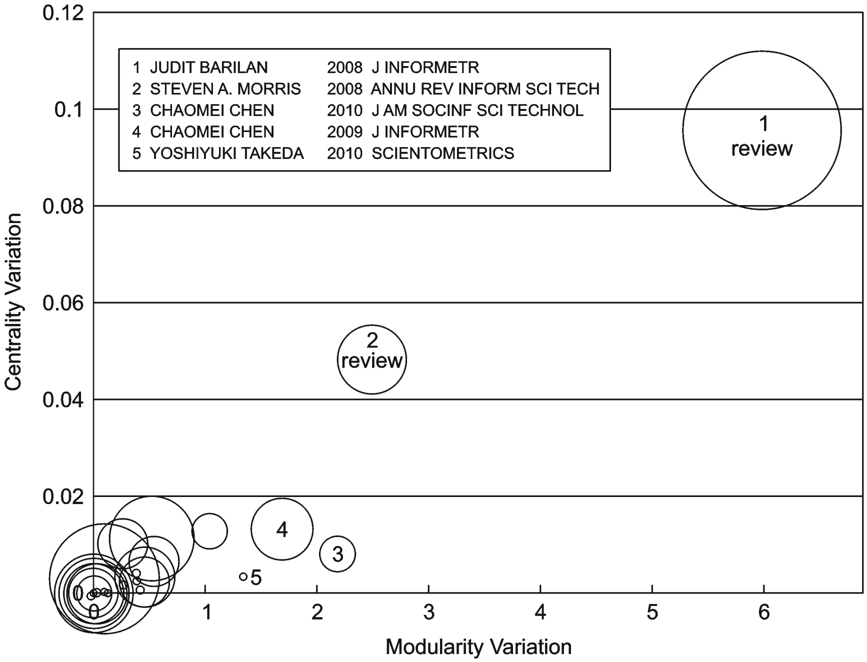


Fig. 8.4 What do the two variation metrics measure? Source: 80 papers citing (Chen, 2006).

rate. The size of a circle is proportional to the number of citations that the corresponding paper received by 2010. As the graph shows, the two papers that were ranked consistently high by both metrics are review papers. What are these metrics really measuring? Is there a reason why would a review paper be ranked high by these variation metrics?

Recall that the modularity variation rate is designed to give higher scores to papers that add unprecedented connections between distinct modules (i.e. clusters) in a network representation of the history of a topic. A paper ranked high by modularity variation should be among the papers that cited references that have not been cited together. A review paper obviously fits into this category. Similarly, a paper with a strong centrality variation rate means that the paper has introduced a considerable shift in the distribution of node centrality, probably by citing references with an unusual pattern or a combination of patterns. Again, a review paper can fit into this category as well. Therefore, the pattern that review papers are ranked high is indeed consistent with the theoretical expectation.

The more interesting cases would be non-review papers that are ranked high. Among the top-5 papers highly ranked by modularity variation, #3∼#5 are original research papers. In other words, these papers are not review papers, but they standout because they add new connections between modules that are previously not connected. According to our theory of discovery, these papers have the potential to transform the research topic. They are likely to become highly cited later on.

Another noteworthy property of the ranking method is that #3 and #5 are papers published in the current year. This shows that the method can identify papers with modularity variation without having to wait for citations to build up. In fact, the method does not rely on any use or evaluative data such as times downloaded, times visited, or times cited. This is one of the distinct advantages of this approach. Users can access to relevant indicators of transformative potential as soon as a paper is published or even when it is submitted for publication.

The results of the UNIANOVA test and the negative binomial regression consistently identified the centrality variation rate is a reliable predictor of citation counts of a paper. This implies that the centrality variation rate as a novelty metric is likely to be meaningful. More thorough tests should be done with larger datasets over a longer period of time to further verify the role of both metrics. Nevertheless, in addition to scientific publications, the same method is applicable to grant proposals, patents, and other sources of information by constructing baseline network representations similarly to the ones we tested with journal papers. In next section, we apply the method in a case study of the pulsars research for the first 10 years of its development.

8.2.4 Case Study: Pulsars

Meadows and O'Connor (1971) described the discovery of pulsars in astronomy as an example of what they believed to be a general tendency in the publication of scientific research. Pulsars were discovered by accident in 1967 while Jocelyn Bell and Antony Hewish were searching for twinkling sources of radio radiation[3]. A distant star twinkles when viewed from the earth. The twinkling of a star is due to the refraction of light through the atmosphere. Scientists build optical telescopes on top of high mountains to reduce such wiggling views. Since radio telescopes are not affected by the Earth's atmosphere, if a radio source is still twinkling, then there must be some other reasons.

One of the most interesting radio sources in space is quasars. Quasars, quasi-stellar, are compact radio sources just like stars but sending out strong radio signals. Scientists believe that quasars may reveal the Universe in its really early stage. Antony Hewish at the Cavendish Laboratory in Cambridge, England, was searching for quasars. He reasoned that radiation from a compact source like a quasar would twinkle more than radiation from a less compact source, i.e. a region, so the more twinkling radio sources are, the more likely they are quasars. Hewish designed a large radio telescope to carry out the search. Jocelyn Bell, a research student of Hewish, was responsible for operating the telescope and visually analyzing the data. Some strange-looking signals drew her attention — the signal always came from the same patch of the sky and it was a series of steady pulses, 1.3 seconds apart. Hewish and Bell considered various explanations for this regular signal coming from outer space. Soon they found another signal pulsing at 1.2 seconds. Now it became less likely that the regular signal comes from aliens; instead, there must be some natural explanation. By January, Hewish and Bell had found four of such pulsing radio sources — pulsars. Their discovery was published on February 24, 1968 in *Nature* (Hewish, Bell, Pilkington, Scott, & Ccollins, 1968).

This discovery marked the beginning of a new research area in astronomy. In 1974 Antony Hewish and Martin Ryle, head of the Cavendish radioastronomy group, were awarded the Nobel Prize in Physics for their pioneering research in radio astrophysics. Hewish was the first astronomer to receive the Nobel Prize in Physics. Jocelyn Bell, who made the initial discovery, was awarded a CBE for her services to astronomy in 1999.

Pulsars are highly magnetized and rapidly spinning neutron stars. They are the remains of stars from supernova explosions in the distant past. Gold's theoretical explanation of *pulsating radio sources* (pulsars) was published on May 25, 1968 in *Nature* (Gold, 1968). Gold predicted that the rotation rate of these neutron stars would gradually decrease as they radiated energy. The prediction was proven to be the case. Pulsars slow down by about one

[3] http://www-outreach.phy.cam.ac.uk/camphy/pulsars/pulsars2_1.htm

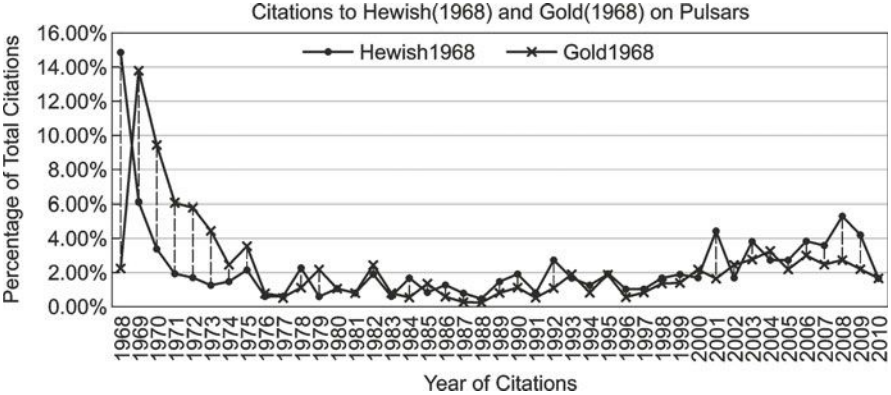


Fig. 8.5 Hewish et al. (1968) has 472 citations as of 2010 July 5th Gold (1968) has 362.

millionth of a second per year. The ratio of a pulsar's current speed to its slow-down rate tells us how old it is.

The terminology was also changed rapidly in the earlier years of pulsar research. The initial term used by the discoverers was *pulsating radio sources.* 72% of the papers used the term were published in 1968 (18 papers). Only 3 papers used the term in 1969 (12%). After 1970, the term was almost completely vanished. In comparison, the term *pulsar* was used in 54 papers published in 1968, in 147 papers in 1969, and 151 papers in 1970. The word *pulsar* is made from *puls*ating st*ar*.

In Meadows and O'Connor's study of the growth of pulsar research, they noticed a distinct initial concentration of pulsar papers in Nature, especially within the first 6 months of the publication of the paper by Hewish et al. Five weeks later, the first two theoretical papers on pulsars appeared, also in *Nature*. Meadows and O'Connor identified this as a general tendency of the birth of a new field: papers in a new growth area tend to concentrate in the same journal as the original discovery paper.

The initial concentration of pulsar papers in Nature attracted 52% of the citations to pulsar papers. The rate dropped to 40% in the first half of 1969. Pulsar papers subsequently appeared in more and more journals. Meadows and O'Connor noticed that it was already evident that journals with pulsar papers appeared to comply with the Bradford law in the period of the first 18 months.

The speed of publication in the initial growth period of the pulsar research area is remarkable. The original discovery paper took two weeks from the reception to its publication in Nature, and one of the first theoretical papers took only five days! The citation half-life of pulsar papers in the first two years of the discovery was 0.7 years, which means 50% of the citations were given to papers published 0.7 year ago. In contrast, the half-life of astronomy papers as a whole in the same time frame was 5.4 years. The short half-life of pulsar papers in the initial years means that researchers had to seek rapid publications; otherwise, their papers would become obsolete before they are even published.

The number of citations per paper in a new area of research also conveys interesting signals. Initially, there are few papers to cite. As relevant papers appear rapidly, the numbers of citations increases rapidly.[4] In the pulsar example, the average number of citations per paper grew from 7.1 in 1968 to 9.9 in 1969. In addition, the rate of self-citation dropped from 15% in 1968 to 10% in 1969 as the research expanded to more research groups.

Another characteristic noticed by Meadows and O'Connor is that initial pulsar papers have a higher number of co-authors on average (2.0 authors per paper) than astronomy as a whole (1.5 authors per paper). More specifically, observational papers had a higher co-author rate of 2.65 than theoretical papers of 1.55.

The rapid increase of papers as both citing and cited papers leads to the expectation that the citation network must become rapidly interwoven. During the initial period of growth, what would be the prominent patterns of citations on the structure of the literature? In terms of a co-citation network, if one paper's citation preserves the citation structure of the new growth and another paper's citation drastically alters the citation structure, can we distinguish which one is likely to be more important than the other, purely based on how their citations influence the existing citation structure?

To illustrate the extent to which macroscopic properties of pulsar papers at earlier stages signal their subsequent impact, we use the co-citation network of 1968 pulsar papers as the baseline reference and measure the degree of the structural change introduced by a new pulsar paper published in the following 5 years. Papers that induce the most profound structural change are regarded having strong brokerage potential that could lead to fundamental changes later on. We expect to see that the brokerage potential measure is an important factor to explain the actual impact observed several years later.

Pulsar papers published in a 10-year window (1968 – 1977) were collected from the Science Citation Index Expanded. Due to the terminology change, we used a topic search for both '*pulsating radio source**' and *pulsar**. Each paper is ranked with a score ΔQ, which is the percentage of network modularity change due to its citation with reference to the network structure formed by prior publications up to the year before, and a score ΔC, which is the relative entropy of the new betweenness centrality distribution over the prior distribution. The citations of these papers are measured globally as of July 6, 2010. The search found 1,048 records. The peak of the subject lasted about 10 years till early 1980s.

The 11-cluster configuration is optimal based on modularity and silhouette scores (Fig. 8.6). Nodes are labeled by sigma scores, which identified Hewish_1968 and Gold_1968 as the most influential discovery papers. Cluster

[4] As a more recent example, the Sloan Digital Sky Survey (SDSS) research doubles its citation and paper counts every 10 months.

Table 8.8 Top-20 cited references by pulsar papers by sigma. The first five papers are published in *Nature* in 1968 and 1969.

Freq	Burst	Centrality	Σ	PageRank	Author	Year	Source	Vol	Page	HalfLife	Cluster
85	21.64	0.11	8.87	2.14	HEWISH A	1968	NATURE	V217	P709	1	0
101	9.14	0.26	8.05	3.44	GOLD T	1968	NATURE	V218	P731	2	5
48	13.15	0.09	3.00	2.36	LYNE AG	1968	NATURE	V218	P326	1	0
61	12.80	0.06	2.21	2.24	RADHAKRISHNANV	1969	NATURE	V221	P443	1	5
45	11.08	0.06	1.93	1.93	LARGE MI	1968	NATURE	V220	P340	1	6
91	8.15	0.08	1.89	2.78	OSTRIKER JP	1969	ASTRO...	V157	P1395	3	5
83	8.61	0.07	1.83	2.76	STURROCK PA	1971	ASTRO...	V164	P529	4	3
44	4.85	0.10	1.62	3.16	MANCHESTER RN	1971	APJS	V23	P283	4	3
81	7.08	0.07	1.58	2.31	PACINIF	1968	NATURE	V219	P145	2	5
59	9.07	0.05	1.55	2.16	GOLDT	1969	NATURE	V221	P25	1	5
50	6.60	0.06	1.49	2.77	LYNE AG	1971	MON N...	V153	P337	4	3
38	7.59	0.05	1.46	2.00	WAMPLER EJ	1969	APJ	V157	L1	2	3
57	24.77	0.02	1.46	1.88	RUDERMAN MA	1975	ASTRO...	V196	P51	1	3
27	8.64	0.04	1.38	1.32	MANCHESTER RN	1974	ASTRO...	V189	L119	2	3
128	2.53	0.12	1.33	3.28	GOLDREICH P	1969	ASTRO...	V157	P869	4	5
41	13.55	0.02	1.29	1.19	COCKE WJ	1969	NATURE	V221	P525	0	8
45	7.15	0.04	1.28	1.75	DRAKE FD	1968	NATURE	V220	P231	2	3
41	12.08	0.02	1.27	1.40	RUDERMAN M	1972	AREV...	V10	P427	3	5
37	13.96	0.02	1.26	1.33	HULSE RA	1975	ASTRO...	V195	L51	1	1
40	4.26	0.05	1.23	2.16	RICKETT BJ	1970	MNRAS	V150	P67	3	3

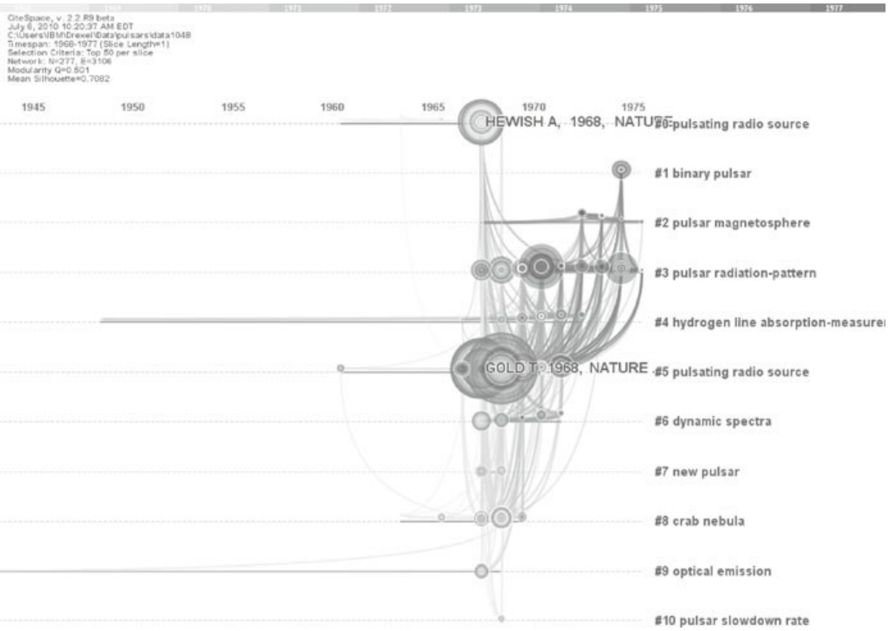


Fig. 8.6 A timeline visualization of 11 clusters of references co-cited by articles on pulsars. Each cluster is labeled by the most prominent terms in citing articles.

#5 is a thread of theoretical papers. Cluster #3 is a thread of papers that validate theoretical predictions.

Table 8.9 shows the result of a full factor model with raw scores of alpha (existing co-citation links) and beta (novel co-citation links) with the change rate of modularity and centrality as the covariant, weighted by year of publication, excluding intercept. The model explains 84% of variance.

Table 8.9 Tests of Between-Subjects Effects[c]
Dependent Variable: Times Cited as of 2010

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Model	5.366E9	115	4.666E7	16.681	.000
$\Delta_{\text{Modularity}}$	8176068.802	1	8176068.802	2.923	.089
$\Delta_{\text{Centrality}}$	8124014.657	1	8124014.657	2.904	.090
beta	1.361E9	28	4.862E7	17.380	.000
alpha	2.029E9	44	4.610E7	16.480	.000
beta * alpha	5.770E7	16	3606329.792	1.289	.205
Error	6.294E8	225	2797501.193		
Total	5.996E9	340			

a. R Squared = .895 (Adjusted R Squared = .841)
b. Computed using alpha = .05
c. Weighted Least Squares Regression – Weighted by Year

Table 8.10 Papers ranked by the modularity variation rate ΔM (ΔC=centrality variation rate, TC=Times Cited, NR=Number of References).

Year	ΔM	ΔC	TC	NR	Article
1970	18.09	0.0558	79	102	HEWISH A, PULSARS, ANNU REV ASTRON ASTROPHYS, V8, P265
1972	17.46	0.0563	9	161	SMITH FG, PULSARS, REP PROGR PHYS, V35, P399
1971	6.88	0.0317	45	200	GINZBURG VL, PULSARS - THEORETICAL CONCEPTS, SOV PHYS USPEKHI-USSR, V14, P83
1972	6.23	0.0181	352	169	RUDERMAN M, PULSARS - STRUCTURE AND DYNAMICS, ANNU REV ASTRON ASTROPHYS, V10, P427
1969	2.57	0.0013	21	25	CHIU HY, RADIO EMISSION FROM MAGNETIC NEUTRON STARS - A POSSIBLE MODEL FOR PULSARS, PHYS REV LETT, V22, P415
1969	2.06	0.0029	0	25	FELDMAN PA, LOW-ENERGY COSMIC RAYS FROM PULSAR FLARES, NATURE, V223, P48
1969	2.06	0.0008	91	13	RADHAKRI.V, EVIDENCE IN SUPPORT OF A ROTATIONAL MODEL FOR PULSAR PSR 0833-45, NATURE, V221, P443
1969	2.06	0.0009	137	17	GUNN JE, MAGNETIC DIPOLE RADIATION FROM PULSARS, NATURE, V221, P454
1972	1.92	0.0085	37	57	MANCHEST.RN, PARAMETERS OF 61 PULSARS, ASTROPHYS LETT COMMUN, V10, P67
1970	1.85	0.0082	6	74	CHIU HY, A REVIEW OF THEORIES OF PULSARS, PUBL ASTRON SOC PAC, V82, P487
1971	1.60	0.0044	136	43	MANCHEST.RN, OBSERVATIONS OF PULSAR POLARIZATION AT 410 AND 1665 MHZ, ASTROPHYS J SUPPL SER, V23, P283
1969	1.54	0.001	9	11	HUNT GC, LINEAR INCREASE IN PERIODICITY OF 13 PULSARS, NATURE, V224, P1005
1972	1.08	0.0104	111	30	BOYNTON PE, OPTICAL TIMING OF CRAB PULSAR NP 0532, ASTROPHYS J, V175, P217
1970	1.05	0.0035	105	46	RANKIN JM, RADIO PULSE SHAPES FLUX DENSITIES AND DISPERSION OF PULSAR NP-0532, ASTROPHYS J, V162, P707

Continued

Year	ΔM	ΔC	TC	NR	Article
1971	0.69	0.0012	25	33	CHIU HY, THEORY OF RADIATION MECHANISMS OF PULSARS .1., ASTROPHYS J, V163, P577
1972	0.67	0.0045	16	22	SHITOV YP, FINE-STRUCTURE OF SPECTRA OF RADIO EMISSION OF PULSARS, ASTRON ZH, V49, P470
1970	0.55	0.0009	397	42	GUNN JE, ON NATURE OF PULSARS .3. ANALYSIS OF OBSERVATIONS, ASTROPHYS J, V160, P979
1971	0.54	0.0033	31	16	HUNT GC, RATE OF CHANGE OF PERIOD OF PULSARS, MON NOTIC ROY ASTRON SOC, V153, P119
1971	0.54	0.0032	22	29	MANCHEST.RN, ROTATION MEASURE AND INTRINSIC ANGLE OF CRAB PULSAR RADIO EMISSION, NATURE-PHYS SCI, V231, P189
1970	0.44	0.0013	10	27	SMITH FG, GENERATION OF RADIO WAVES IN PULSARS, NATURE, V228, P913

Fig. 8.7 depicts pulsars papers based on their corresponding transformative potentials measured by the modularity variation rate (x-axis) and the

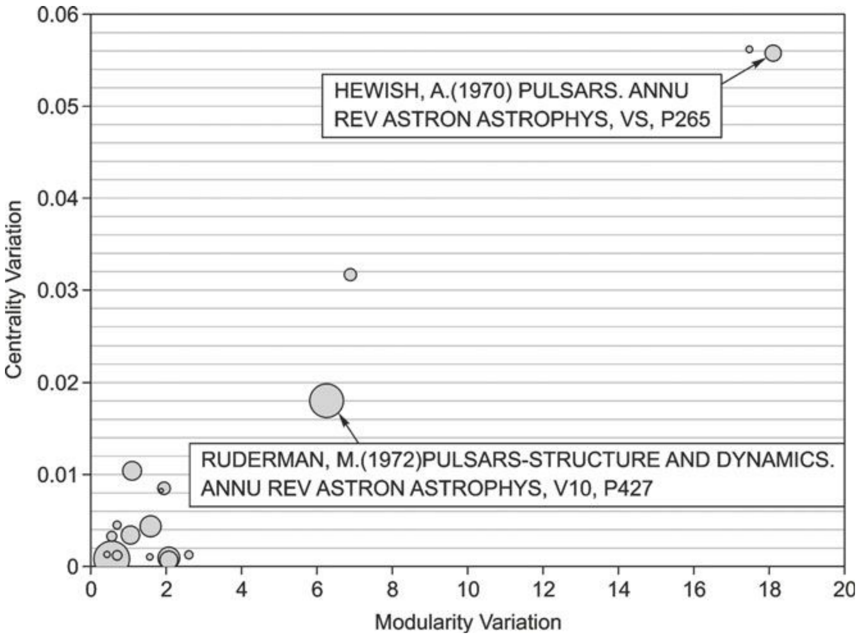


Fig. 8.7 Early review papers on pulsars.

centrality variation rate (y-axis). According to our theory of discovery, contributions with the most transformative potential are expected to appear near the top-right corner, whereas more conventional research should be distributed near to the lower-left region. The two papers labeled in the chart are both review papers. The one at the top-right corner was written by Hewish, the Nobel laureate for his work in this area.

In the remaining of this chapter, we will discuss how the same method can be applied to grant proposals and awarded projects. For proposals and award abstracts, there is no readily available citation data. We demonstrate that the novelty detection method based on structural variation can be applied to network representations of words and noun phrases.

8.3 Portfolio Evaluation

The following example is based on a report we produced for the NSF CISE/SBE Advisory Committee's Subcommittee on Discovery in Research Portfolios. The analysis was done between October 2009 and October 2010. The subcommittee was charged with identifying and demonstrating techniques and tools that characterize a specific set of proposal and award portfolios. The subcommittee was asked to identify tools and approaches that are most effective in deriving knowledge from the data provided, i.e. most robust in terms of permitting program officers to visualize, interact, and understand the knowledge derived from the data. Subcommittee members were asked to apply their research tools to structure, analyze, visualize, and interact with data sets provided by the NSF.

Grant proposals submitted to the NSF consist of a number of components, including a cover page, a one-page project summary, a project description up to 15 pages, a list of references, 2-page biographies of investigators, and budget information. The abstracts of awarded projects are publically available on the NSF's website. In the following analysis, we distinguish two sources of data: the publically available award abstracts and the proposal dataset that was made available to the members of the subcommittee for a limited period of time. All the results discussed in this book regarding the proposal dataset have been approved by a specific clearance procedure, which was in place to safeguard the privacy and security of the proposal dataset.

We focused on questions at two levels. At the individual proposal level, the main questions are: What is a proposal about in a nutshell? How does one proposal differ from other proposals in terms of their nutshell representations? At the portfolio level, the questions focus on characteristics of a group of proposals. What are the computational indicators that may differentiate awarded and declined proposals? What are the indicators that may identify transformative proposals in a portfolio?

8.3.1 Identifying the Core Information of a Proposal

We make no assumption about the structure or content of text documents. We process them as unstructured text. We hypothesis that the core information of a 15-page proposal is probably equivalent to text that is more than the one-page summary but much less than 15 pages of text. If this is true, then we can use a considerably less amount of text to represent the essence of a 15-page document. We can also expect that the shorter representation is likely to be more coherent than the original full-length document.

The full-length project description of a proposal is divided into a series of passages of text (known as segments) so that each passage corresponds to an underlying theme or topic. Then a network of these passages is constructed based on how similar these passages are to one another. Passages that pull together the most of passages are chosen to represent the core information of the proposal. Figure 8.8 shows the interface of our prototype that visualizes the resultant text segments and corresponding text in one of our own proposals. The plot at the bottom of the figure shows the similarity between neighboring texts (in sliding windows of text).

The first step is to divide a full-length project description into a series of passages of text. The internal cohesiveness of text within a passage is higher than between passages. The process is known as *text segmentation*. The basic assumption is that a text document usually represents a series of subtopics and it should be possible to detect the boundary of a subtopic based on the change of text similarities. We adopt Marti Hearst's text segmentation algorithm, which gives encouraging results in terms of its flexibility. Setting optimal parameters for the algorithm is a crucial step, but currently techniques for computationally optimizing the configuration do not appear to be available. We choose to provide interactive visualization tools so that users can intuitively inspect the influence of various configurations and optimize the configuration accordingly.

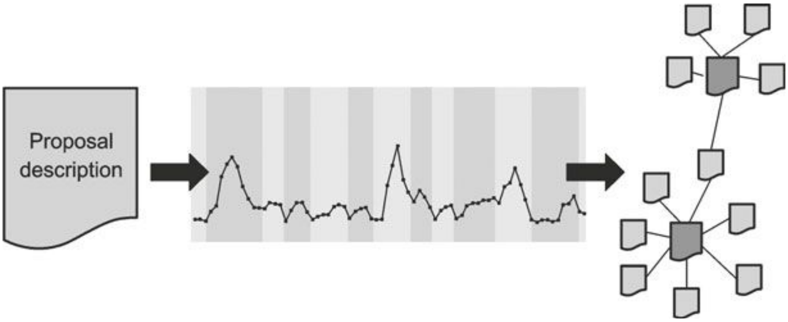
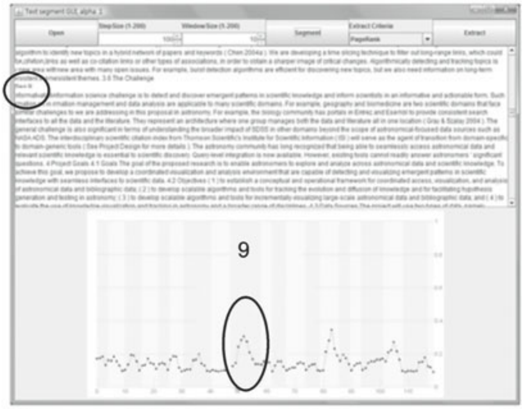


Fig. 8.8 The procedure for identifying core passages of a full-length document.

Hearst's algorithm for text segmentation detects the subtopic shift patterns from lexical differences of a set of n-gram (n is in the range of 10$\sim$200) text strings. We implemented this method and added a visualization of the lexical differences to help users to find the optimal parameter combinations. Two parameters, *windowsize* and *stepsize*, are crucial in configuring the text segmentation algorithm. Windowsize is the number of tokens, mainly terms excluding stop words, in a token-sequence, and stepsize is the number of token-sequence in a block that are used for block-block similarity comparison. The similarity score between two blocks is computed by a normalized inner product: given two blocks, b_1 and b_2, each with *windowsize* token sequence, where b_1=\{token-sequence $_{i-k}$,. . . , token-sequence$_i$\} and b_2=\{token-sequence $_{i+1}$,. . . , token-sequence$_{i+k+1}$\}. Our test with a small number of texts found that *windowsize*=100 and *stepsize*=20 gave optimal results in narrative texts similar to the NSF proposals.

The goal of the second step is to select the most representative segment(s) as the core information of a proposal. Once the text segments are identified in step 1, many existing techniques from the information retrieval community and machine-learning community in particular are available to compute the similarity between any two segments, including vector space models, latent semantic indexing, probabilistic models, and topic models. A network of segments can be constructed based on such segment-segment similarities. Choosing the most representative block of text becomes choosing the segment(s) that have the most significant positions in the network. Our assumption is that the most representative text should be among the ones that have strong centrality properties such as PageRank, degree centrality, and betweenness centrality. In other words, a central topic is expected to be highly connected with other topics in the same proposal. Metrics such as PageRank are able to rank segments in the order of such centrality. One may choose one or multiple top-ranked segments to represent the proposal for any subsequent text analysis, clustering, or visualization.



algorithm to identify new topics in a hybrid network of papers and keywords (Chen 2004a). We are developing a time slicing technique to filter out long-range links, which could be citation links as well as co-citation links or other types of associations, in order to obtain a sharper image of critical changes. Algorithmically detecting and tracking topics is a new area with new area with many open issues. For example, burst detection algorithms are efficient for discovering new topics, but we also need information on long-term persistent [illegible] themes. 3.6 The Challenge

Rank: 9

[illegible] information science challenge is to detect and discover emergent patterns in scientific knowledge and inform scientists in an informative and actionable form. Such [illegible] information management and data analysis are applicable to many scientific domains. For example, geography and biomedicine are two scientific domains that face similar challenges to we are addressing in this proposal in astronomy. For example, the biology community has portals in Entrez and Esembl to provide consistent search interfaces to all the data and the literature. They represent an architecture where one group manages both the data and literature all in one location (Gray & Szalay 2004). The general challenge is also significant in terms of understanding the broader impact of SDSS in other domains beyond the scope of astronomical-focused data sources such as NASA ADS. The interdisciplinary scientific citation index from Thomson Scientific's Institute for Scientific Information (ISI) will serve as the agent of transition from domain-specific to domain-generic tools (See Project Design for more details). The astronomy community has long recognized that being able to seamlessly access astronomical data and relevant scientific knowledge is essential to scientific discovery. Query-level integration is now available. However, existing tools cannot readily answer astronomers ' significant questions. 4 Project Goals 4.1 Goals The goal of the proposed research is to enable astronomers to explore and analyze across astronomical data and scientific knowledge. To achieve this goal, we propose to develop a coordinated visualization and analysis environment that are capable of detecting and visualizing emergent patterns in scientific knowledge with seamless interfaces to scientific data. 4.2 Objectives (1) to establish a conceptual and operational framework for coordinated access, visualization, and analysis of astronomical data and bibliographic data; (2) to develop scalable algorithms and tools for tracking the evolution and diffusion of knowledge and for facilitating hypothesis generation and testing in astronomy; (3) to develop scalable algorithms and tools for incrementally visualizing large-scale astronomical data and bibliographic data; and (4) to evaluate the use of knowledge visualization and tracking in astronomy and a broader range of disciplines. 4.3 Data Sources The project will use two types of data, namely

Segments of A Project Description(our own proposal)

Fig. 8.9 A prototype that assists users to find the core information of a proposal.

Our preliminary study evaluated these candidate ranking metrics against our own proposals and found that segments selected by PageRank are more meaningful than others considered. However, a larger scale evaluation is needed and it will be useful if test data sets can be constructed and made available to us for evaluation.

8.3.2 Information Extraction

Information extraction is another essential task for processing and analyzing unstructured text. The goal is to filter words and phrases that are closely related to the point and to the case that the investigator intends to make. Natural language processing (NLP) techniques are used for this purpose. We particularly focus on noun phrase extraction based on the assumption and the general consensus that noun phrases are more meaningful and interpretable than single words.

The input for this step can be any stream of text, structured or unstructured. This step takes the core segments identified in a proposal as the input and generates a list of noun phrases found in the core segments. The core information identification operates as the first layer of filter to collect text data for subsequent processing. Noun phrase extraction is the second layer of filter.

The incoming stream of text is first tagged by NLP techniques known as Part-of-Speech (POS) tagging, which explicitly marks the type of each and every word in the text so that the targeted words, i.e. noun phrases, can be recognized by extraction algorithms. The extraction of noun phrases follows heuristics, ranging from simple to complex ones. The output of the process is a list of noun phrases found in the representative text of a proposal and how many times they appear.

Noun phrases consist of multiple words with a noun as the final word (known as the head noun). For example, the blackhole in supermassive blackhole is the head noun. The word *term* may refer to single or multiple words, but not necessarily contain nouns. Sometimes terms are also referred to as n-grams, with n as the number of component words. Noun phrases are considered in general better lexical units to represent concepts than words and terms because noun phrases tend to be more meaningful and self-contained.

In order to extract noun phrases, the first step is to tag the type of each word in a text passage, including nouns, verbs, and adjectives. This step is called Part of Speech (POS) tagging. Natural language processing (NLP) tools are available to perform POS tagging and process tagged text, for example, GATE. Since NLP tools tend to be built with particular training text, the quality of tagging varies from target datasets. However, we found that using regular-expression is the most flexible, customizable, and extensible approach. We experimented with several types of noun phrases in terms of the number of nouns because a general form of a noun phrase is word-word-word-noun and it is possible the words are also nouns themselves, for example, word-word-noun-noun as in *rapidly increased climate change*. We allow users to filter noun phrases in terms of the number of nouns in a phrase. We tested analytical and statistical results across noun phrases with different word counts.

8.3.3 Detecting Hot Topics

Hot topics are defined in terms of the frequency of noun phrases found in project descriptions, project summaries, or other sources of text. Generally speaking, high-frequency noun phrases are regarded as indicators of a possible hot topic. The most valuable information of a hot topic is therefore when it becomes hot and how long it lasts as a hot topic with reference to other topics occurring at the same time.

There are many techniques for detecting the timing of a hot topic. In this project, we adopt Kleinberg's burst detection algorithm and detect when the frequency of a noun phrase starts to jump and how long it remains high. We use these two measures in subsequent survival analysis to differentiate awarded and declined proposals.

Burst detection determines whether the frequency of an observed event is substantially increased over time with respect to other events of the same type. The types of events are generic, including the appearance of a keyword in newspapers over a period of 12 months and the citations to a particular reference in papers published in the past 10 years. The data mining and knowledge discovery community has developed several burst detection algorithms. We adopt Kleinberg's algorithm for its flexibility.

Two major measures of noun phrase burst are considered in our work: the waiting time to burst and the duration of burst. The waiting time to burst is how long the time elapses between the initial appearance of a noun phrase in a set of proposals and the time when a burst is statistically detected. The duration of burst is the time elapses between the beginning of the burst until either the burst drops or the end of the timeframe of the analysis. These measures are used in the subsequent survival analysis that aims to differentiate awarded and declined proposals. These measures are domain independent and do not require additional semantic-related input.

8.3.4 Identifying Potentially Transformative Proposals

We envisage that the nature of transformative research should be detectable along two of the computationally observable dimensions: *synthesis distance* and *structural divergence*. The synthesis distance characterizes a particular scientific contribution in terms of the conceptual distance between component topics it synthesizes and integrates. The larger the distance, the harder it can be conceived but also potentially of higher novelty. The structural divergence measures the extent to which a particular scientific contribution departures from the state of the art. As illustrated in Fig. 8.10, ground-breaking ideas are expected to have a distant synthesis distance and a large structural divergence.

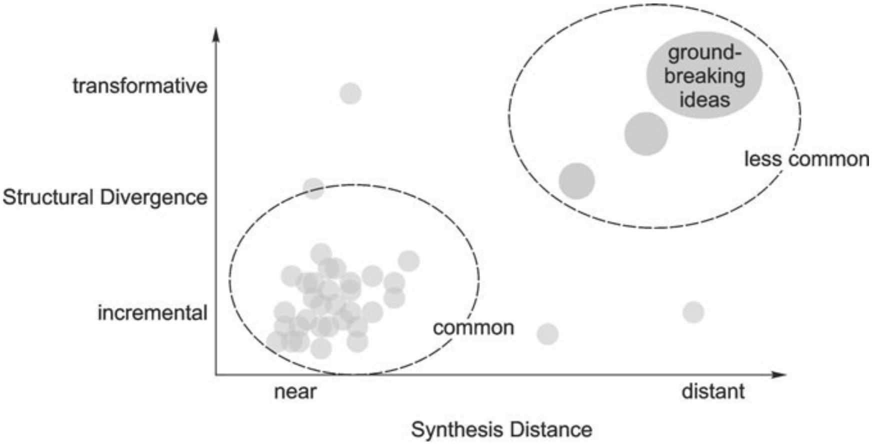


Fig. 8.10 An illustration of the distribution of transformative research along two dimensions.

Fig. 8.11 shows the preliminary results of an analysis of the publically available abstracts of awards of the NSF SciSIP program. The size of each award represents the amount awarded. The position of an award is determined by the two metrics along synthesis distance and structural divergence. The details of each of the four awards annotated in Fig. 8.11 are shown in Table 8.11, including the PI's name, the year of award, and the title of the project.

Table 8.11 Awards labeled in Fig. 8.12.

Lee Fleming (2008) DAT: Creating a Patent Collaboration Network Database to Examine the Social Production of Knowledge
Feldman Maryann (2008) State Science Policies: Modeling Their Origins Nature Fit and Effects on Local Universities
Martin Ribarsky (2009) DAT: A Visual Analytics Approach to Science and Innovation Policy
Philip Shapira (2008) MOD Measurement and Analysis of Highly Creative Research in the US and Europe

We also conducted the second preliminary test with 200 proposals (100 awarded and 100 declined) randomly sampled from 7,345 proposals of an NSF program. The core information of each proposal is represented by the

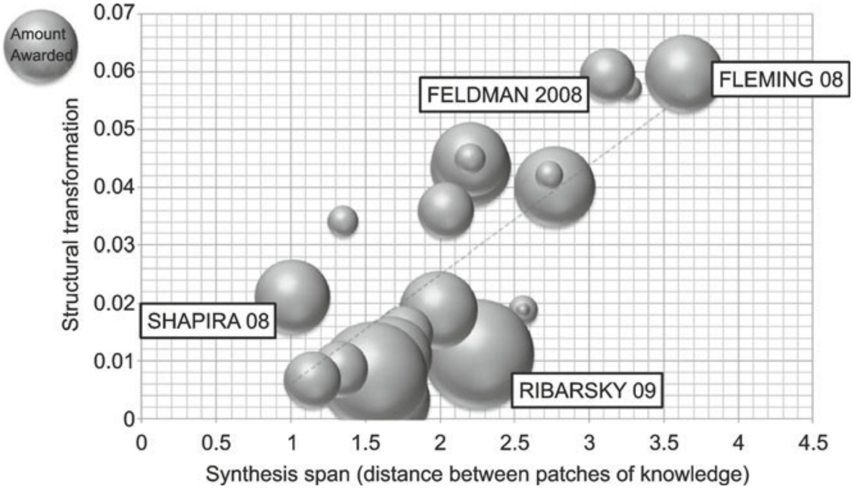


Fig. 8.11 Transformative potentials of awards. Data Source: Publically available NSF award abstracts of the SciSIP Program.

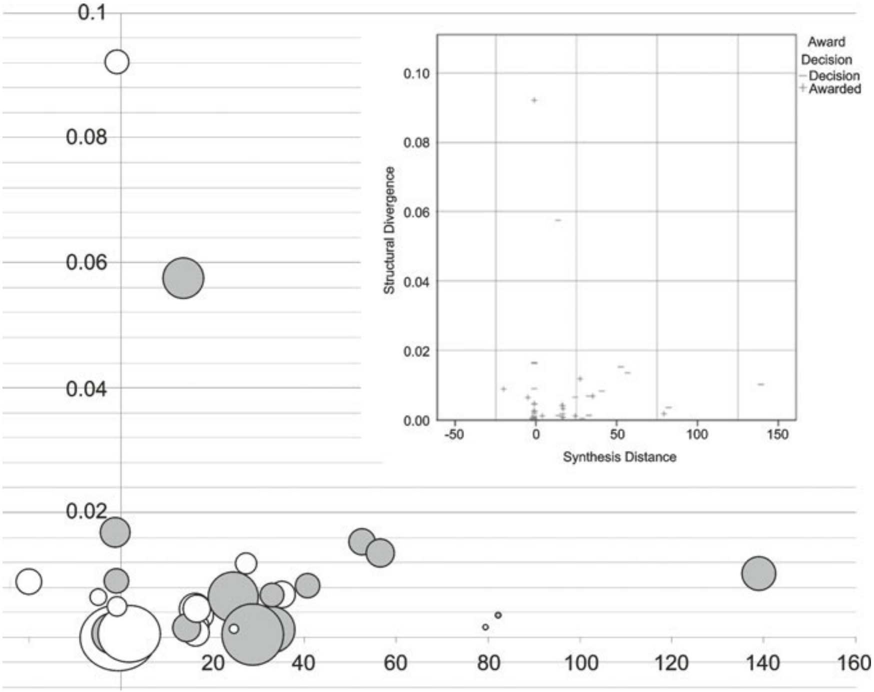


Fig. 8.12 Transformative potentials of proposals (+/circle: awarded; -/grey declined). Size=the amount requested. Data Source: 200 randomly sampled proposals.

top-3 ranked core segments chosen by PageRank. A new version of CiteSpace was used to extract noun phrases of 1 $\sim$ 4 nouns and generate the two measures for each proposal. A total of 141 proposals (70.5%) were found to have positive readings on the chart (see Fig. 8.12).

8.4 Summary

Core information identification techniques such as text segmentation and network-based ranking are relatively mature and scalable. However, the optimal configuration of these techniques requires extensive and in-depth evaluation. Noun phrase extraction has a considerable overhead, which is the need for POS tagging. POS tagging is time consuming, but for a given dataset it only needs to be done once and subsequent processing can repeatedly use the result of the POS tagging. Therefore, it is advisable to minimize the interdependence among the various components in this process. Burst detection is also relatively time consuming, but much faster than POS tagging. It is also a one-time only need if the same dataset is used.

Techniques for identifying the transformative potential of proposals are still at a very early stage of research. Preliminary results are very encouraging and pointing to many potentially fruitful directions to pursue. We continue the research and development in terms of its theoretical underpinning and technical advances.

References

Chen, C. (2003). Mapping scientific frontiers: The quest for knowledge visualization. London: Springer.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57(3), 359-377.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics, 3(3), 191-209.

Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The Structure and dynamics of co-citation clusters: A multiple-perspective Co-Citation Analysis. Journal of the American Society for Information Science and Technology, 61(7), 1386-1409.

Gold, T. (1968). Rotating neutron stars as origin of pulsating radio sources. Nature, 218(5143), 731-732.

Härynen, M. (2007). Breakthrough research: Funding for high-risk research at the Academy of Finland. Helsinki: The Academy of Finland.

Hewish, A., Bell, S.J., Pilkington, J.D.H., Scott, P.F., & Collins, R.A. (1968). Observation of a rapidly pulsating radio source. Nature, 217(5130), 709-713.

Lee, Y.G., Lee, J.D., Song, Y.I., & Lee, S.J. (2007). An in-depth empirical analysis of patent citation counts using zero-inflated count data model: The case of KIST. Scientometrics, 70(1), 27-39.

Lipinski, C., & Hopkins, A. (2004). Navigating chemical space for biology and medicine. Nature, 432(7019), 855-861.

Lokker, C., & Walter, S.D. (2010). Prediction of citation counts: a comparison of results from alternative statistical models. Retrieved Oct 15, 2010, 2010, from http://www.bmj.com/content/336/7645/655/reply

Meadows, A.J., & O'Connor, J.G. (1971). Bibliographical statistics as a guide to growth points in science. Science Studies, 1(1), 95-99.

NSF. (2007, September 25). Important notice No. 130: Transformative research. http://www.nsf.gov/pubs/2007/in130/in130.txt. Accessed 14 Aug 2010.

Price, D.D. (1965). Networks of scientific papers. Science, 149, 510-515.

Swanson, D.R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine, (30), 7-18.

Takeda, Y., & Kajikawa, Y. (2010). Tracking modularity in citation networks. Scientometrics, 83(3), 783-792.

Chapter 9 The Way Ahead

In this final chapter, we first summarize the key points in the previous chapters and how they are connected or may influence each other. Then we identify a few theoretical and practical issues that need to be dealt with in future studies and applications.

9.1 The Gathering Storm

Before the Gathering Storm, a primary source of concern for individual scientists, their institutions, and public funding agencies is the decreasing level of public funds available for the increasing demands of research funding. In this context, as funding agencies become more and more stringent in selecting projects to fund, researchers respond to the declining successful rate by increasing the number of their submissions. At the same time, funding agencies and institutions increasingly find themselves in a position to address accountability issues: what are the projects you decided to fund but you shouldn't, what are the projects you didn't find but you should, and how do you justify the way the public fund is allocated? On the one hand, taxpayers rightfully expect that their money should be used to fund research that is likely to benefit the society. On the other hand, it is also known that applied science and technological innovations are built on basic research and that the societal implications of basic research are not always clear, in fact, almost always not clear. We cannot demand eggs while rejecting the responsibility for feeding the hens. How do we resolve this dilemma with the limited resources?

The Gathering Storm and related debates are not only pressing on these issues even harder but also drawing our attention to more profound questions. What is the key for maintaining and sharpening the competitive edge of a nation? To address these questions, we narrow down from the funding crunch to the nature of creativity and the role of individual scientists in sustaining the leading position of a nation in science and technology. The Yuasa phenomenon is a macroscopic pattern, but it may have microscopic explanations. The average age of scientific productivity offers one possible

C. Chen, *Turning Points*

explanation of why the world center of scientific activity may move or stay. It hints the potential role of scientific creativity, but like many macroscopic-focused approaches, theories along this line do not offer much of a constructive guidance; there is not much we can do about our age. What are the actions beyond the age that we can take? What is it in the ways of our thinking that we can consciously enhance so that not only can we come up with more original and creative ideas, but also find creative solutions to hard problems?

9.2 Creative Thinking

In Chapter 2 we reviewed what we, collectively, know about creative thinking. In contrast to the widespread belief in the serendipitous nature of creativity in general and scientific creativity in particular, we paid special attention to generic mechanisms of creative thinking and problem solving. Starting with the consensus of the need for divergent thinking and thinking outside the box, we reviewed mechanisms that could be used to carry out an open-minded search for new ideas and possible solutions and difficulties that one is likely to encounter. Two categories of profound challenges emerge. One is similar to finding needles in a haystack. The other is how to keep our mind open.

The modern versions of finding needles in a haystack intensify the challenge in opposite directions. Not only has the size of the haystack become enormously large, in the magnitude of 10^6 and higher, the traces of needles have become increasingly invasive. One is no longer certain whether a needle found is really the needle that one is looking for. A key question is how an intellectual pathway is formed in such a vast space. How do we decide the direction of our next move? Darwinian thinking has been influential in terms of the blind variation and selective retention paradigm.

On the surface, the second category has more to do with the blind variation aspect of the paradigm than many have realized, but selective retention can be profoundly influenced by the perspective that one happens to align with. Due to the wide variety of human cognitive biases and weaknesses, we often select variations from a narrow band of possible ones framed by a biased perspective. In other words, our blind variation is often not 'blind' enough, and not open-minded enough. The literature we reviewed underlines the fact that it is intrinsically hard for us to switch from one perspective to another, for example, from the types of Gestalt switches that define scientific revolutions to the lack of imagination of what aliens may look like.

Several general strategies to break away from existing thinking systematically are found in the literature, including ones that reference to the existing thinking and ones that do not. The examples of the former include thinking of something different with reference to the existing thinking by thinking of the opposite of existing thinking, by integrating existing thinking from different sources, and by breaking up and recombining existing thinking. What these strategies have in common is that they broaden the horizon of our new thinking in a systematic way and bring alternatives into the equation. Janusian thinking makes it clear that we should take the opposite of what we are interested into account, which would probably make it easier to justify why public funding should support antimatter research. In fact, modern techniques for matrix decomposition such as non-negative decomposition are able to identify multiple dimensions from input data. Thinking of different dimensions across is equivalent to applying Janusian thinking to the multi-dimensional problem. Moving from one dimension to another is challenging for us because it often involves a change of perspective, i.e., a Gestalt switch.

TRIZ can be seen as taking one step further by explicitly framing a problem as part of a contradiction and the goal is to resolve the contradiction. In other words, being able to make a Gestalt switch is no longer the goal; instead, the goal now is to be able to see all contradictory aspects of the same phenomenon simultaneously and by doing so the previously perceived contradictions will disappear. To be fair, examples given for Janusian thinking already imply the motivation to consider seemingly contradictory phenomena simultaneously. The nature of creative thinking is the ability to come up with theories that can consistently explain contradictions, or crises in Thomas Kuhn's structure of scientific revolutions.

9.3 Biases and Pitfalls

Chapter 3 is concerned with biases and pitfalls that one may encounter or has to cope with along the way of searching for creative ideas and recognizing them. Our mental models, our perspectives, and working theories are not only simplified but also biased representations of the world. The same evidence can be used by different parties for their own purposes. As shown in the examples of 911 terrorist attacks and Iraq WMDs, data do not speak for themselves, theories and models do!

Rejecting future Nobel Prize worthy papers raises more questions about how the transformative potential of research is recognized at various stages of research, from the early grant proposals, intermediate publications, to widely recognized impacts on society. Experts who propose new research projects can be overly optimistic, whereas experts who serve the role of peer reviewers may have legitimate reasons to reject premature ideas and poorly articulated research plans. On the other hand, from a social and community point of view, peer review experts technically do have a conflict of interest — they are competitive peers.

Conflicts of interest aside, how hard is it to recognize the potential of a research topic? In Chapter 4, we have looked at both hindsight and foresight of scientific breakthroughs. Project Hindsight and TRACES looked retrospectively into the past and provided many lessons. If Project Hindsight focused more closely on the selective retention phrase of creativity, TRACES emphasized the blind variation phrase more. Lessons learned from TRACES show that technical innovations are preceded by years of mission-oriented critical events, which are in turn preceded by even longer periods of non-mission research. It is particularly hard to justify the potential values of non-mission research to the society.

Early warning signs might be available and detectable as a complex system is about to experience a phase transition, or a transformative change, although some changes take place without any early signs or clues at all. Early warning signs serve as navigational cues for navigators in the vast space of the unknown. Although the optimal foraging theory is not introduced until Chapter 5, the presence or absence of early signs will make qualitative differences as they will tip the balance between perceived risks and rewards. The change in the ratio of perceived risks and rewards will change the diffusion and feedback within the system. The self-reinforced feedback will cascade the initial impact and accelerate the transformation of the system.

The reflection on the history and findings of foresight activities is to reinforce the theme of the chapter that human cognition is biased at both individual and collective scales. Recent assessments of the accuracy of forecasts made by earlier foresight surveys indicated that experts tend to be over optimistic. Although researchers offer explanations why experts make overly optimistic predictions, it is still not clear for the practical implications of the over optimistic tendency on the foresight seeking activities as a whole. Soliciting stakeholders' judgments on attractiveness of research topics is a move in expanding the scope of social contract between science and society. The attractiveness ratings from stakeholders provide the best justification of the social values or at least the potential of social values of research topics. The downside is that such attractiveness rating schemes are intrinsically limited to mission-oriented and development phrases of scientific and technological breakthroughs as demonstrated by the TRACES study; they will not be reliable and even feasible for judging non-mission science.

Assessments of foresight activities so far have generally missed the broader issue: to what extent the high-impact scientific breakthroughs were ever identified as the priority areas by foresight-seeking activities? If the NSF, the DoD, or the Office of Science and Technology were to commission another Project Hindsight, TRACES, or Hindsight on Foresight today, how many transformative discoveries achieved today were given priorities based on experts' consensus 20, 30, or 50 years ago? What were the signs that made experts to pick their feasibility ratings right? Who were the visionary users back then to identify the attractiveness of transformative discoveries before their conception?

9.4 Foraging

Chapter 5 introduces the critical decision making criteria in a broadly defined foraging process — profitability. The foraging metaphor provides an explanatory framework for scientific discovery in particular and creativity in general. The basic assumption is that it is more likely to find creative ideas across the boundaries of currently disparate patches of known knowledge than in the middle of a well-studied patch. We tend to be interested by theories, claims, and interpretations that challenge what we know and yet appear to be rational and logical. We like to be surprise a little but not too much. This tendency echoes creative thinking strategies such as Janusian and TRIZ. A recently published study of patents has reached a similar conclusion that boundary spanning is a good indicator of the quality of patents as long as the gap is not too small and not too large either. The second assumption is that scientists can recognize early signs of transformative work and they leave trails in the literature. Building on the two assumptions, we introduce an explanatory and computational theory of transformative discovery. According to the theory, potentially transformative contributions are detectable in terms of some computational properties of structural and temporal patterns. We do not expect that this theory will capture all sorts of transformative discoveries, nor do we expect that contributions meet these criteria will all turn out to be truly transformative. Even if we can only identify a small number of truly transformative research in this way — our case studies of Nobel Prize winning discoveries have demonstrated that this is possible — then it will justify the value of this new approach for its low cost and highly repeatable on demand nature. The additional potential of the approach is that it can be used as a tool to broaden our horizon by identifying novel connections with reference to our own mental model as soon as they appear in the literature.

The sticky effect revealed in the case study of gene targeting reinforces the appropriateness of the foraging metaphor. Scientists are maximizing the returns of their intellectual foraging by capitalizing on a rich patch of opportunities. Scientists pick their priority areas to maximize their intellectual returns. When financial constraints become an issue, scientists will re-assess the risk to reward ratio and act accordingly. Since foresight surveys typically focus on a 20$\sim$30 year timeframe, research is needed to improve our understanding of how science advances and how priority areas evolve from one generation of scientists to next.

9.5 Knowledge Domain Analysis

Chapter 6 starts to present a series of quantitative approaches to the study of scientific change. Starting from single snapshots of the structure of a scientific field, we introduce progressive knowledge domain analysis that studies how a topic, a field, or a discipline evolves over time. A time series of snapshots are stitched together into a panoramic view of the history of the domain. Intellectual turning points, or pivotal points of paradigm shifts, are identified and presented. Critical paths of evolution are visualized.

The second part of Chapter 6 moves the state of the art further. A critical missing link in the traditional scientometric studies is evidence-based sensemaking and analytics. In an analog of a flock of flying birds in the sky and their shadows on the ground, traditional studies would study and interpret the shadows of flying birds. The focus of citation analysis, for example, is more often on the value of highly cited papers than on the new contexts in which they play their roles. Similarly, the focus of co-citation analysis tends to focus on clusters of co-cited papers. When analysts label co-citation clusters, they often derive their labels from the target of citations, i.e., the shadows, rather than the source of citations.

The multiple-perspective co-citation analysis is designed to shift the focus from the shadows to the flying birds so that analysts, students, and scientists can explore the new contexts in which earlier published papers are cited and what these new citing papers have in common.

Chapter 6 shows what and how we may learn from the knowledge represented in the literature. Knowing what we know collectively is a necessary step towards learning how transformative discoveries have been made and recognized in the past and what mechanisms and indicators we might develop to enhance our creativity.

9.6 Text Analysis

Chapter 7 focuses on temporal patterns and variations in text. The first part of the chapter gives an example of distinguishing conflicting opinions, namely positive and negative reviews made by Amazon customers on a bestseller *the Da Vinci Code.* The second part introduces a method designed for extracting patterns from unstructured text without relying on any predefined ontology or taxonomy. The method is designed based on an observation that the more important a topic is to an author, the more language variations are likely to be used to describe the topic. This phenomenon can be found not only at a document level, but also at a cultural level. For example, the Chinese language has a much richer set of vocabulary to describe relatives than languages such as English do. In Chinese, there is one specific word for an elder brother and for a younger brother, whereas in English, it needs a combination of two words to express the same meaning. This particularly refined vocabulary on this topic reflects the significance of these meanings in Chinese culture. By the same token, in *the Origin of Species* Darwin wrote many things about *species*, *forms*, *plants*, and *differences.* These patterns emerged from text expressed in a natural language.

The design of the method is in fact more ambitious. Its ultimate goal is to provide a baseline representation of the current scientific knowledge. Concepts can be naturalistically identified from natural language passages. Relations and predicates can be identified in the basic form of subject — verb — object. The known and the unknown can be represented as assertions, claims, and hypotheses associated with available evidence and a level of uncertainty. Newly proposed scientific ideas can be compared against the master representation of the knowledge and their novelty can be derived and inferred.

The third part is about detecting the burstness of topical patterns. Burst detection aims to identify the intensity and duration of an elevated level of activities. For example, citation burst is defined a period of time in which citations to a paper exceed a given threshold or a probabilistically defined transition rate. The burst of the occurrences of a word can be similarly defined as a period of time in which the frequency of the word is exceedingly high with regards to other words.

The fourth part on survival analysis is relatively new. Although survival analysis as a statistical method is widely used, the combination of burst detection and survival analysis is novel. Survival analysis enables us to compare two or more groups in terms of their temporal patterns. In particular, survival analysis allows us to address questions such as between highly cited and less highly cited groups of publications on a topic, which group is more likely to contain topics that are bursted sooner? Which group is more likely to find topics that sustain their bursts longer?

9.7 Transformative Potential

Chapter 8 introduces the concept of transformative potential of scientific research embodied in a scientific publication, a patented idea, a grant proposal, or an awarded project. The measurement of transformative potential is theory driven. It is particularly derived from the explanatory and computational theory of transformative discovery. Simply speaking, transformative potential is measured along two dimensions in examples given in this chapter, but it can be measured by other dimensions.

The central idea in measuring the transformative potential is that structural variations provide early signs. According to our theory of transformative discovery, ideas that introduce a greater degree of structural variation are more likely to have the potential to transform the knowledge structure than those that alter the existing structure to a less extent. We have developed two metrics along this line of reasoning. The synthesis span metric measures the degree of structural variation in terms of the distance between the existing structure and a new structure. The structure can be a network representation or a probabilistic distribution of multiple topics and citation clusters. In other words, the synthesis span indicates the amount of boundary spanning implied by the research in question.

The structural divergence measures the overall change between the old and the new structures in terms of the centrality measures of individual entities. This metric assumes a network representation. Intuitively speaking, a high score of this metric will identify contributions that cause a significant shift of centers of concentrations in the existing network. If we were to apply this to a network of world scientific activities, it would track the shift of the world center of scientific activities.

These metrics are theory driven. One way to assess their validity is to see to what extent they are good predictors of how soon and how well an associated research embodiment is recognized. For scholarly publications, citations are generally regarded as a reasonable indicator of impact, at least how much attention peer scientists paid to cited publications.

The list of suspects of good predictors of citation has been getting longer and longer over the years. Reviews and survey papers are known to attract a big share of citations. Papers written by many co-authors from prestigious universities are suspected to be citation attractors. The number of references cited by a paper is also considered as a possible factor. There are many models and many independent variables are involved.

An intriguing criterion of a good theory is its coherence. A coherent theory provides a simpler explanation of the same phenomenon than previous theories. We have demonstrated with our preliminary results that our explanatory and computational theory of transformative discovery offers a much simplified explanation of why and how scientific papers are cited. The underlying boundary spanning mechanisms provide a consistent explanation of why review papers tend to be cited more, why papers citing more references tend to be cited more, and why papers with a diverse group of co-authors tend to be cited more.

Obviously more work is still to be done. Nevertheless, the initial results are very encouraging indeed. Not only can we summarize the state of the art as often as we wish, but also access to alternative means of identifying the transformative potential of newly emerged ideas and even what-if and other speculations. What are the possible roles that these new methods can play in enhancing the creativity of individuals, in recognizing the potential of transformative research proposed by others, or in providing alternative methods of assessing feasibility and attractiveness?

9.8 Recommendations

Several lessons learned are particularly worth noting along with a few recommendations for individual researchers, students, and science policy makers and funding agencies.

First, the self-assessment and the courage to face long-term challenges in opening debates such as the Gathering Storm are essential to sustain the leading position of a nation in science and technology as well as in economic, political, and cultural sectors.

Second, foresight-seeking activities need longitudinal follow-up assessments. Retrospective assessments should pay close attention not only to how priority areas identified earlier evolved, but also to scientific breakthroughs emerged in the same timeframe as a whole — regardless whether they were once identified as strategic priority areas. More TRACES-styled studies should be commissioned by funding agencies independently and jointly so that critical events at various stages of the development of transformative science and technology can be closely tracked, understood, and disseminated.

Third, biases and pitfalls in human cognition and decision making should be studied systematically in connection with generic and specific mechanisms for divergent thinking and problem solving.

Fourth, the foraging and brokerage mechanism-based theory of transformative discovery is valuable because it is able to reduce a large number of possible factors to fewer and more fundamental ones. There are certainly types of discoveries that are beyond the reach of the theory. It is therefore important to identify other types of mechanisms that could explain other types of discoveries.

Fifth, quantitative and visual analytic methods become increasingly capable of tracking the evolution of the intellectual dynamics of a scientific domain. More theories should be developed to guide the design and use of such tools.

The most important message of the book is twofold:

- Creativity often arises from carefully considering conflicting conceptualizations.
- Creativity can be cultivated and enhanced with a better understanding of generic mechanisms and potential early signs as well as an improved awareness of biases, pitfalls, and cognitive traps.

Creativity is the ability and willingness to embrace the unknown with an open mind!

Index

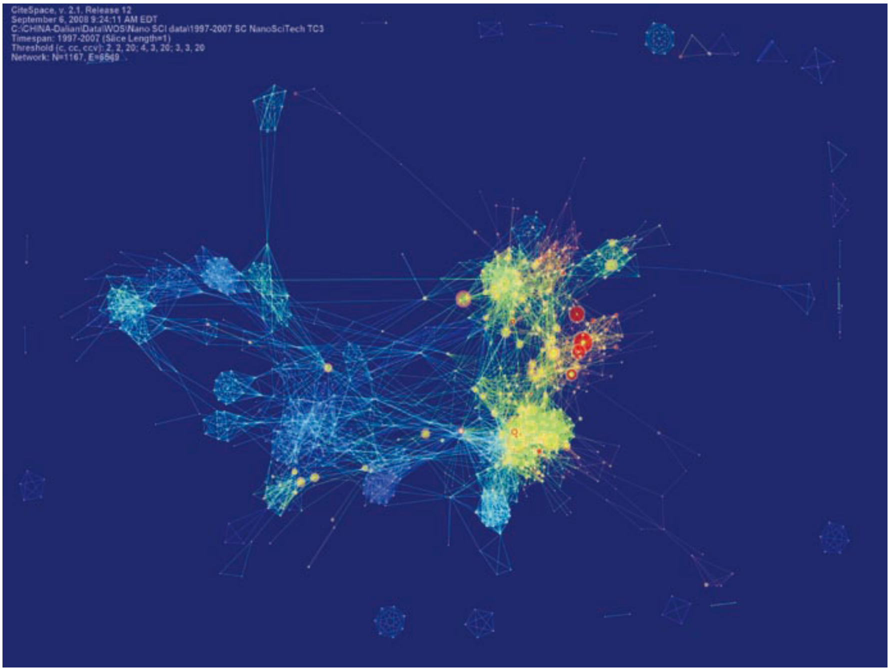

CiteSpace, v. 2.1, Release 12
September 6, 2008 9:24:11 AM EDT
C:\CHINA-Dalian\Data\WOS\Nano SCI data\1997-2007 SC NanoSciTech TC3
Timespan: 1997-2007 (Slice Length=1)
Threshold (c, cc, ccv): 2, 2, 20; 4, 3, 20; 3, 3, 20
Network: N=1167, E=6569

Fig. 1.2 The intellectual trails of the field of nanoscience between 1997 and 2007.

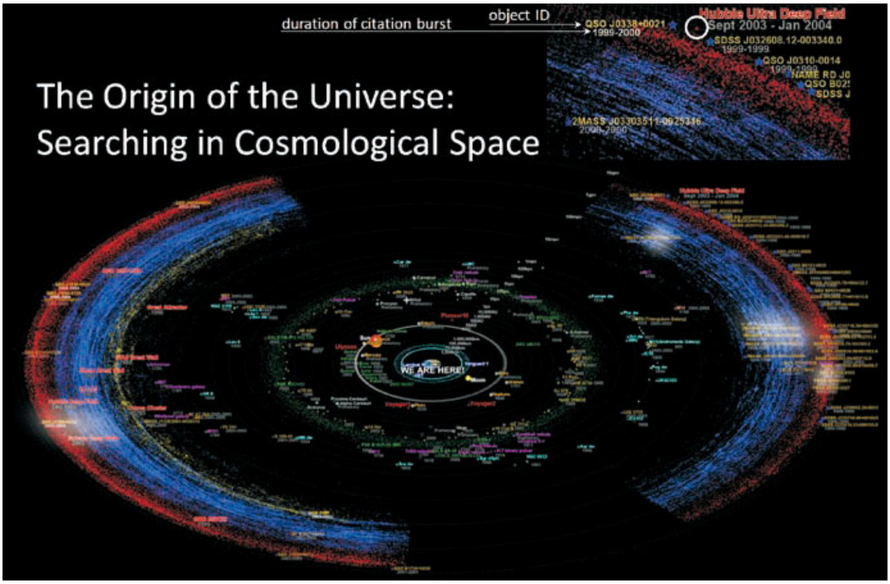


Fig. 1.3 A map of the Universe with overlays of discoveries and astronomical objects associated with bursts of citations. The close-up view of the Hubble Ultra Deep Field is shown at the upper-right corner (circled).

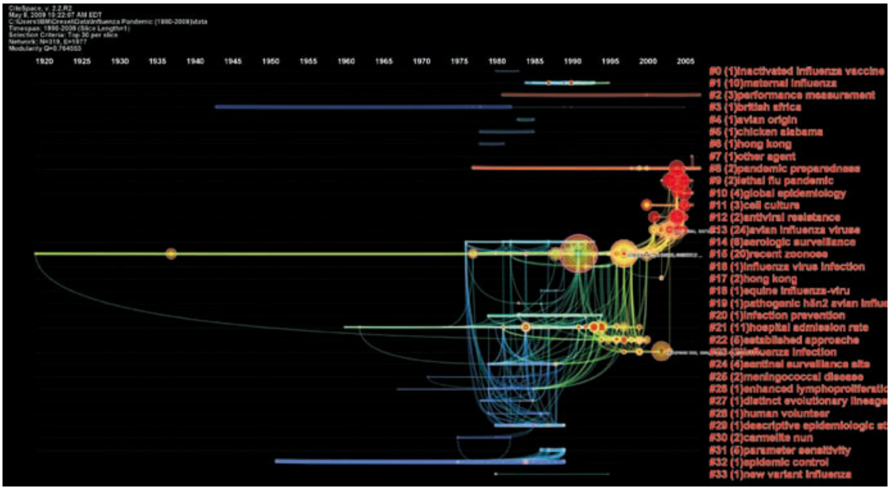


Fig. 1.4 A timeline visualization of the state of the art in research related to influenza and pandemics as of May 8th, 2009.

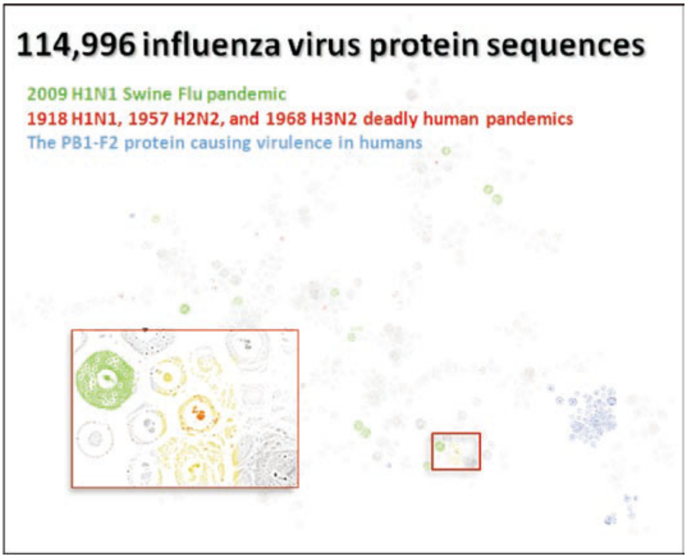


Fig. 1.5 114,996 influenza virus protein sequences. Source: (Pellegrino & Chen, 2011)

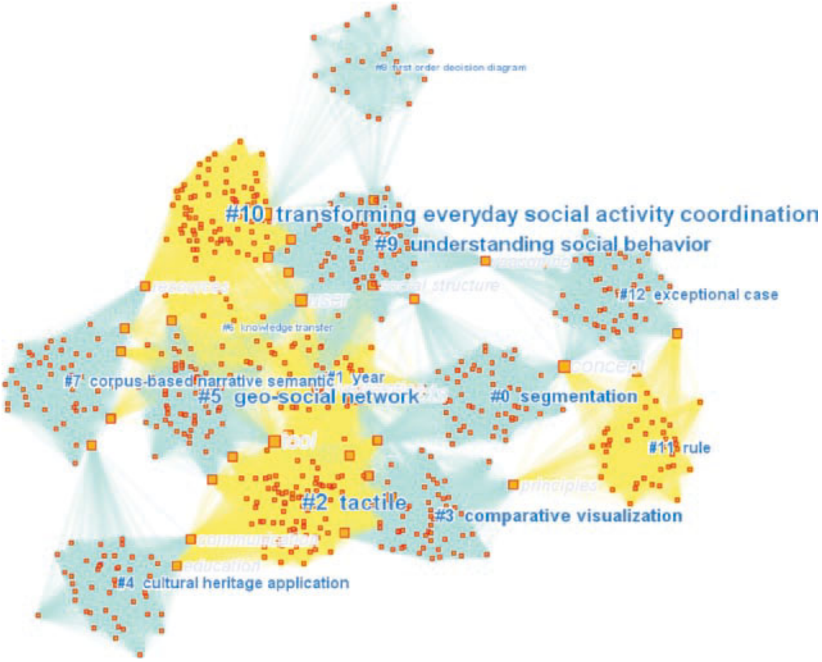

#8 first order decision diagram
#10 transforming everyday social activity coordination
#9 understanding social behavior
reasoning
resources
social structure
user
#12 exceptional case
#6 knowledge transfer
concept
#7 corpus-based narrative semantic
#1 year
#5 geo-social network
#0 segmentation
tool
#11 rule
principles
#2 tactile
#3 comparative visualization
communication
education
#4 cultural heritage application

Fig. 1.6 A network of 682 co-occurring terms generated from 63 NSF IIS EAGER projects awarded in 2009 (cyan) and 2010 (yellow). Q = 0.8565, Mean silhouette = 0.9397. Links = 22347.

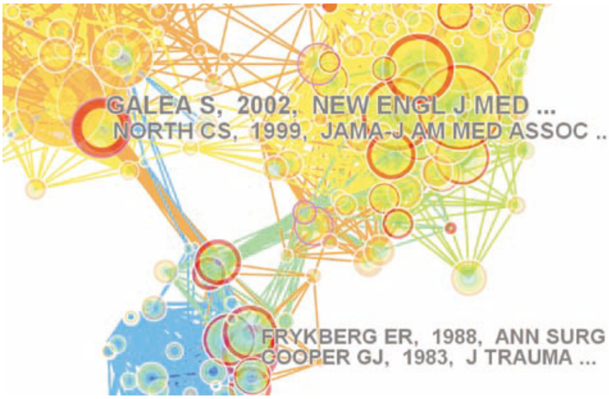

GALEA S, 2002, NEW ENGL J MED ...
NORTH CS, 1999, JAMA-J AM MED ASSOC ..
FRYKBERG ER, 1988, ANN SURG
COOPER GJ, 1983, J TRAUMA ...

Fig. 5.1 The three clusters of co-cited papers can be seen as three patches of information. All three patches are about terrorism research. Prominently labeled papers in each patch offer information scent of the patch. Colors of patches, indicating the time of a connection, provide a scent of freshness. The sizes of citation rings provide a scent of citation popularity. Source: (Chen, 2008).

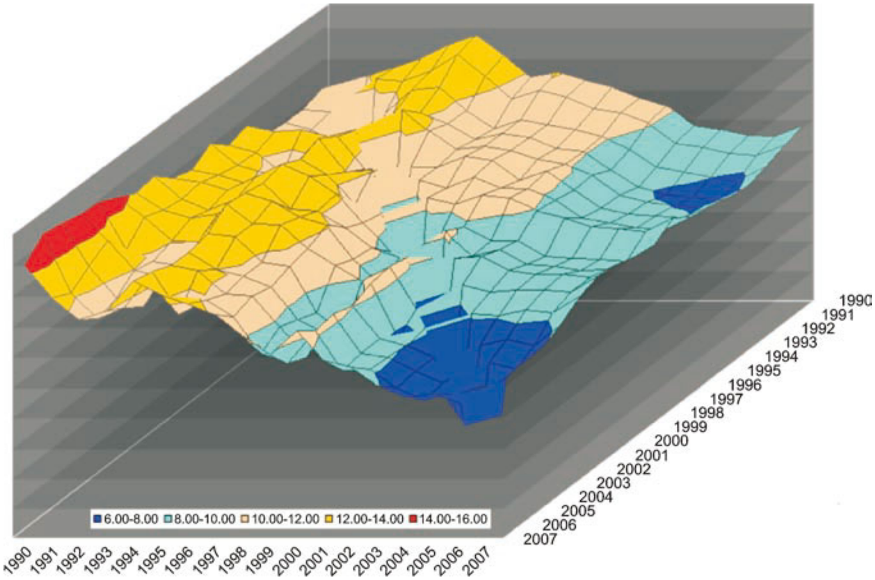
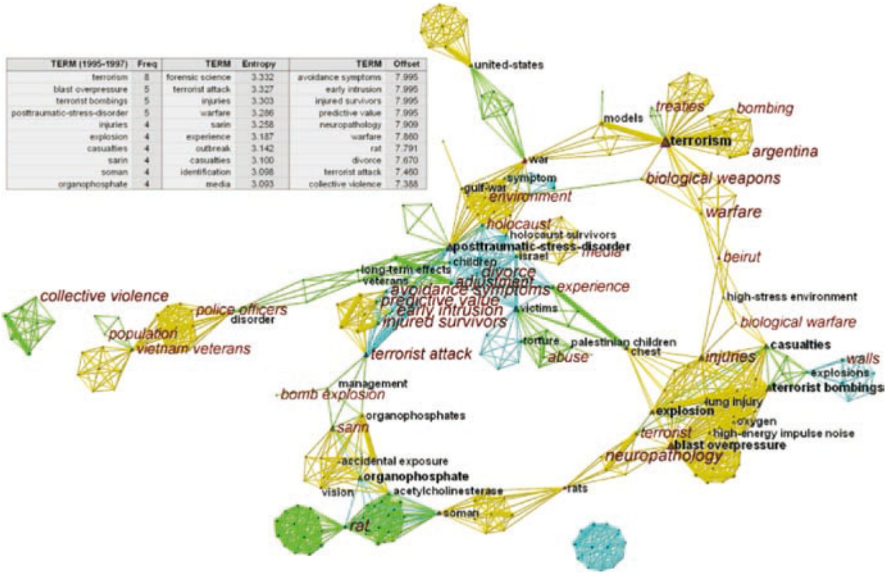


Fig. 5.3 Symmetric relative entropy matrix shows the divergence between the overall use of terms across different years. The recent few years are most similar to each other. The boundaries between areas in different colors indicate significant changes of underlying topics. Source: (Chen, 2008).



TERM (1995-1997)	Freq	TERM	Entropy	TERM	Offset
terrorism	8	forensic science	3.332	avoidance symptoms	7.995
blast overpressure	5	terrorist attack	3.327	early intrusion	7.995
terrorist bombings	5	injuries	3.303	injured survivors	7.995
posttraumatic-stress-disorder	5	warfare	3.286	predictive value	7.995
injuries	4	sarin	3.258	neuropathology	7.909
explosion	4	experience	3.187	warfare	7.860
casualties	4	outbreak	3.142	rat	7.791
sarin	4	casualties	3.100	divorce	7.670
soman	4	identification	3.098	terrorist attack	7.460
organophosphate	4	media	3.093	collective violence	7.388

Fig. 5.4 A network of keywords in the terrorism research literature (1995 – 1997). High-frequency terms are shown in black, whereas outlier terms identified by informational bias are shown in dark red. Source: (Chen, 2008).

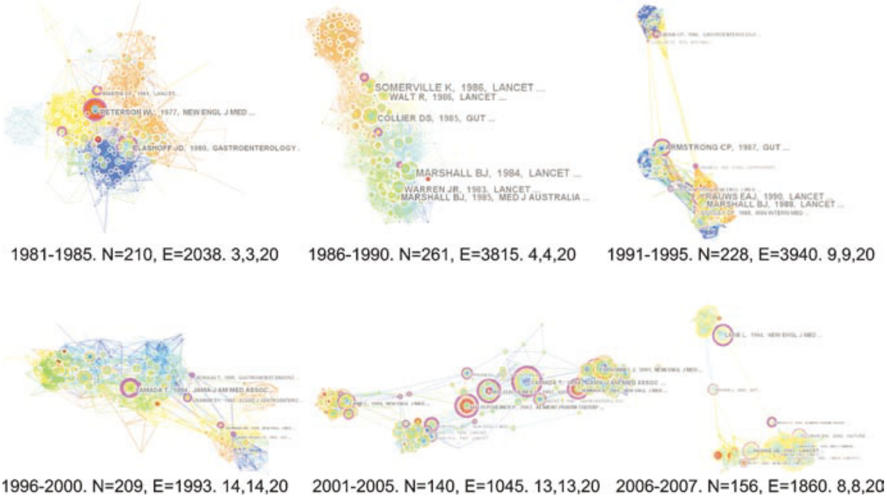

PETERSON WL, 1977, NEW ENGL J MED ...
ELASHOFF JD, 1980, GASTROENTEROLOGY ...
SOMERVILLE K, 1986, LANCET ...
WALT R, 1986, LANCET ...
COLLIER DS, 1985, GUT ...
MARSHALL BJ, 1984, LANCET ...
WARREN JR, 1983, LANCET ...
MARSHALL BJ, 1985, MED J AUSTRALIA ...
ARMSTRONG CP, 1987, GUT ...
RAUWS EAJ, 1990, LANCET ...
MARSHALL BJ, 1988, LANCET ...
YAMADA T, 1994, JAMA-J AM MED ASSOC ...
1981-1985. N=210, E=2038. 3,3,20
1986-1990. N=261, E=3815. 4,4,20
1991-1995. N=228, E=3940. 9,9,20
1996-2000. N=209, E=1993. 14,14,20
2001-2005. N=140, E=1045. 13,13,20
2006-2007. N=156, E=1860. 8,8,20

Fig. 5.6 A co-citation network of references on peptic ulcer research (1980 – 1990). Source: (Chen, Chen, Horowitz, Hou, Liu, & Pellegrino, 2009).

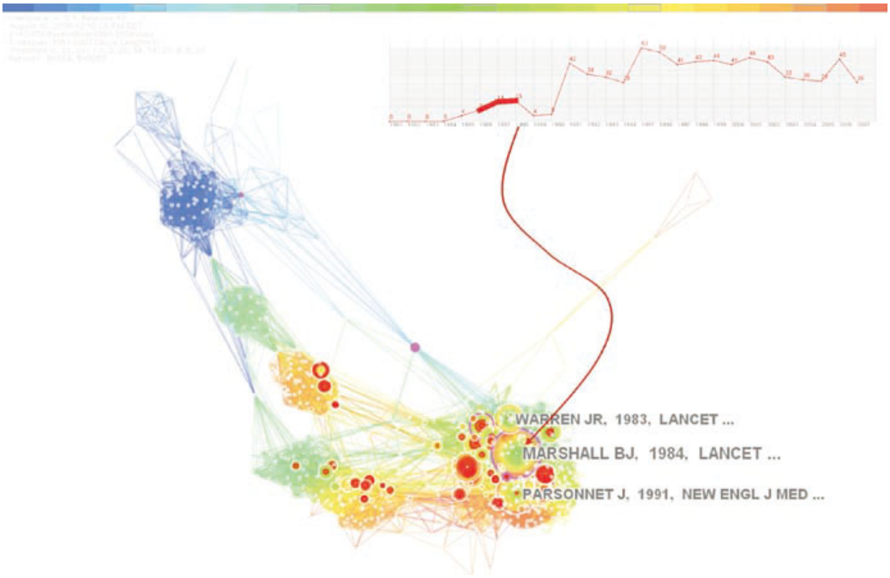

WARREN JR, 1983, LANCET ...
MARSHALL BJ, 1984, LANCET ...
PARSONNET J, 1991, NEW ENGL J MED ...

Fig. 5.7 A co-citation network of references cited between 1981 and 2007 in peptic ulcer research. Source: (Chen et al., 2009).

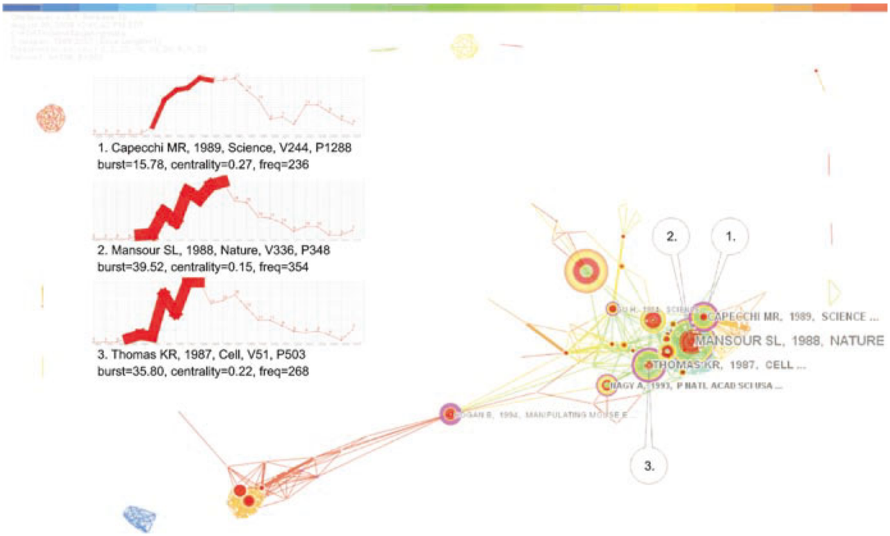


Fig. 5.8 A co-citation network of references cited between 1985 and 2007 in gene targeting research. References with the strongest betweenness centrality scores are labeled. The burst periods of their citations are shown as the thickened curves in the three diagrams to the left. Source: (Chen et al., 2009).

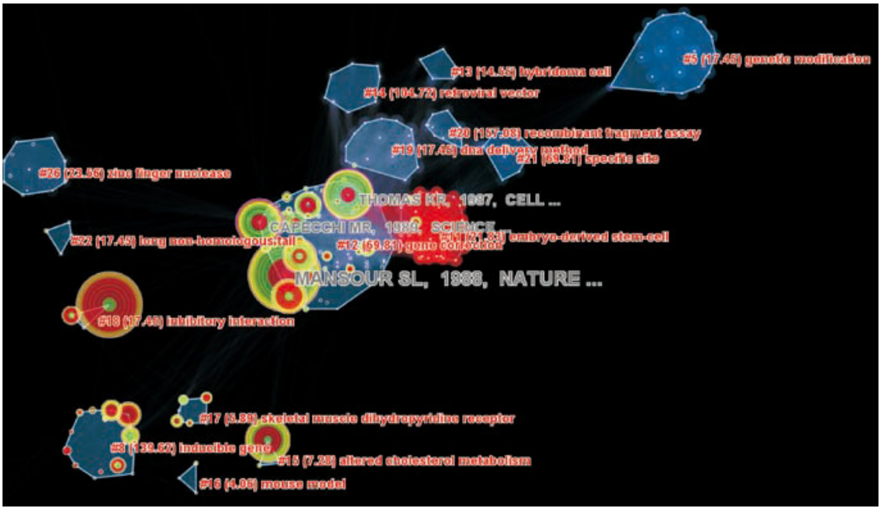


Fig. 5.10 A diffusion map of gene targeting research between 1985 and 2007. Selection criteria are at least 15 citations for citing articles and top 30 cited articles per time slice. Polygons represent clusters of co-cited papers. Each cluster is labeled by title phrases selected from papers citing the cluster. Red lines depict co-citations made in the current year. The concentrations of red lines track the context in which co-citation clusters are referenced. Source: (Chen et al., 2009).

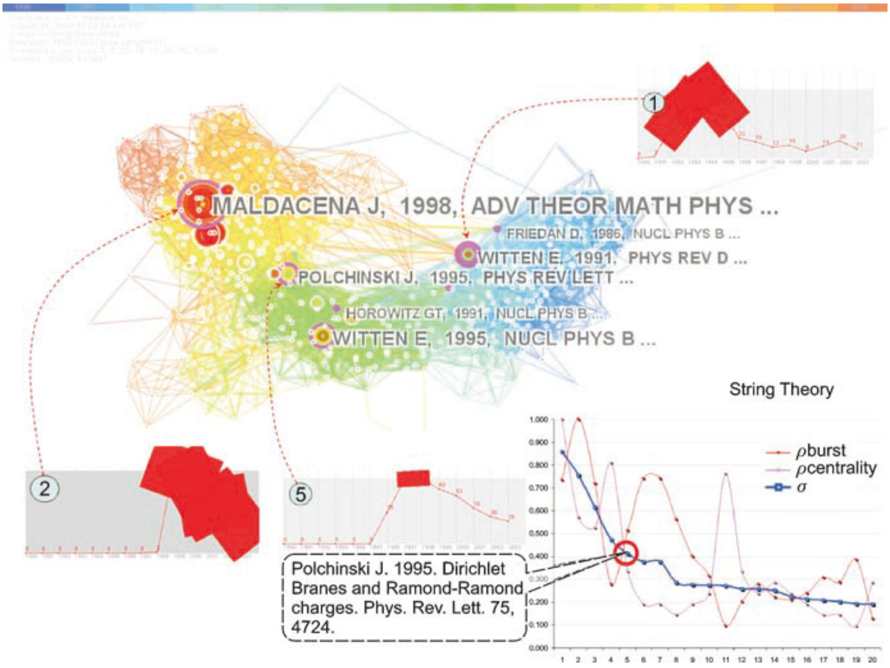


Fig. 5.11 A co-citation network of references cited between 1990 and 2003 in string theory. Polchinski-1995 marked the beginning of the second string theory revolution. Maldacena-1998 is highly transformative and brokerage link between string theory and particle theories. The three embedded plots show the burst periods of citations of Witten-1991, Maldacena-1998, and Polchinski-1995. Source: (Chen et al., 2009).

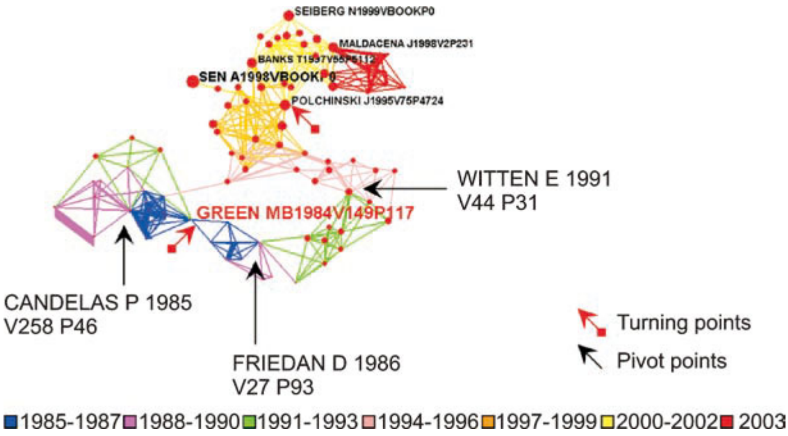


Fig. 6.4 Turning points in superstring research. Source: (Chen, 2004).

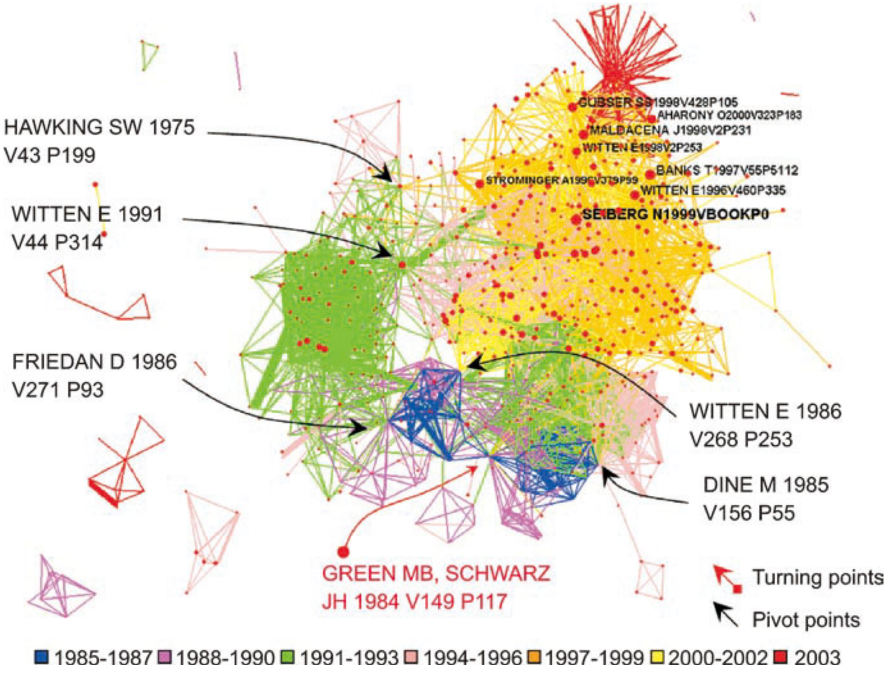


Fig. 6.5 A network of 624 co-cited references. Source: (Chen, 2004).

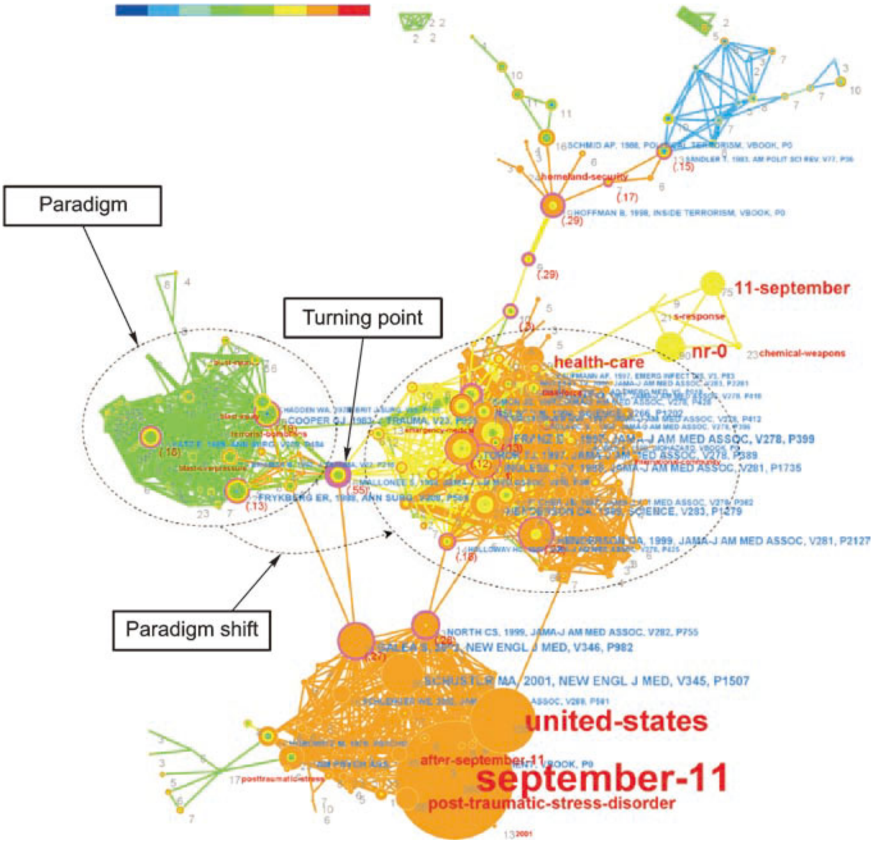

Paradigm
Turning point
Paradigm shift
SCHMID AP, 1988, POLITICAL TERRORISM, VBOOK, P0
SANDLER T, 1983, AM POLIT SCI REV, V77, P36
homeland-security
(.15)
(.17)
HOFFMAN B, 1998, INSIDE TERRORISM, VBOOK, P0
(.29)
(.29)
11-september
s-response
nr-0
chemical-weapons
health-care
HADDEN WA, 1978, BRIT J SURG
COOPER GJ, 1983, J TRAUMA, V23, P955
FRANZ DR, 1997, JAMA-J AM MED ASSOC, V278, P399
TOROK TJ, 1997, JAMA-J AM MED ASSOC, V278, P389
INGLESBY TV, 1999, JAMA-J AM MED ASSOC, V281, P1735
MALLONEE S, 1996, JAMA-J AM MED ASSOC
FRYKBERG ER, 1988, ANN SURG, V208, P569
HENDERSON DA, 1999, SCIENCE, V283, P1279
HENDERSON DA, 1999, JAMA-J AM MED ASSOC, V281, P2127
HOLLOWAY HC, 1997, JAMA-J AM MED ASSOC, V278, P425
(.18)
(.13)
(.12)
(.13)
(.55)
(.18)
NORTH CS, 1999, JAMA-J AM MED ASSOC, V282, P755
GALEA S, 2002, NEW ENGL J MED, V346, P982
(.28)
(.27)
SCHUSTER MA, 2001, NEW ENGL J MED, V345, P1507
SCHLENGER WE, 2002, JAMA-J AM MED ASSOC, V288, P581
united-states
after-september-11
september-11
post-traumatic-stress-disorder
posttraumatic-stress
2001

Fig. 6.6 Major areas in terrorism research. Source: (Chen, 2006).

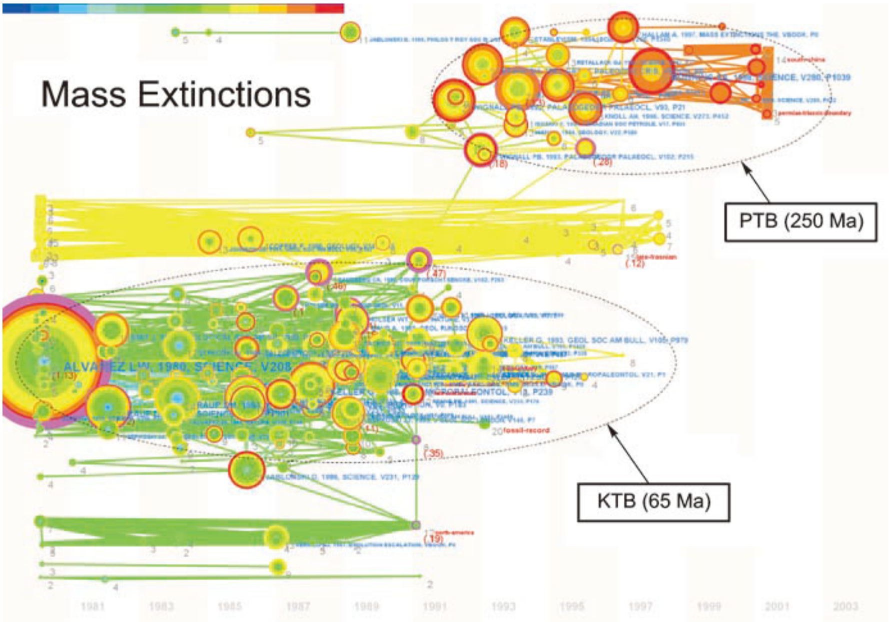


Fig. 6.7 Trends in mass extinctions research. Source: (Chen, 2006).

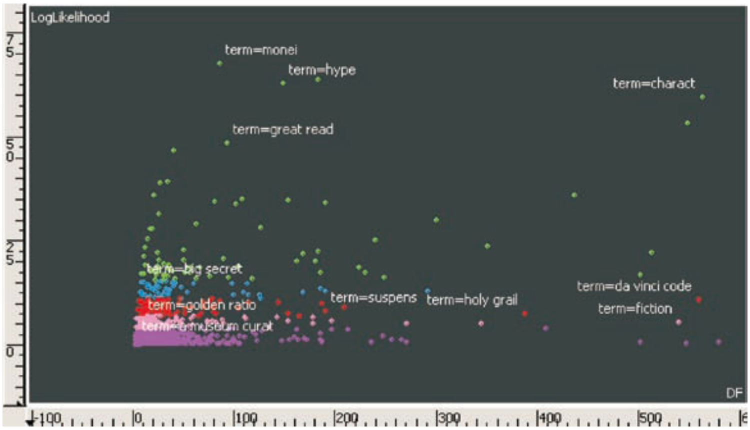


Fig. 7.5 Distributions of selected terms. The colors of dots indicate the statistical significance level of the corresponding terms, namely green (< 0.001), blue (p=0.001), red (=0.01), and pink(=0.5). Source: (Chen et al., 2006).