CiteSpace: A Practical Guide for Mapping Scientific Literature

Book · December 2016

CITATIONS
12

READS
3,685

1 author:

Chaomei Chen
College of Computing and Informatics, Drexel University
390 PUBLICATIONS **8,150** CITATIONS

SEE PROFILE

Dr. Chaomei Chen is a Professor of Informatics in the College of Computing and Informatics at Drexel University. His research ranges from information visualization and visual analytics to knowledge domain visualization, mapping scientific frontiers, and the study of scientific discovery and creativity. He is the Editor-in-Chief of Information Visualization and the Special Chief Editor of Frontiers in Research Metrics and Analytics. He is the author of The Fitness of Information: Quantitative Assessments of Critical Information (Wiley, 2014), Turning Points: The Nature of Creativity (Springer, 2011), Information Visualization: Beyond the Horizon (Springer 2004, 2006), and Mapping Scientific Frontiers: The Quest for Knowledge Visualization (Springer 2003, 2013).

This book is a practical guide not only on how to optimize the use of CiteSpace for visual analytic studies of scientific literature but also on the design rationale and how to interpret various patterns as part of a visual thinking process. The book is written with a minimum amount of jargon. It users everyday language to explain how we may learn valuable insights on the shoulders of giants – how generations of scholars and domain experts have addressed the topics of our interest.

"CiteSpace provides a multifaceted tool suite to analyze and visualize patterns and relationships based on Web of Science search sets (with capabilities to handle other database sources as well). It is a major aid in social network analyses, providing advanced capabilities to treat data properties (e.g., to normalize) and rich science mapping options. CiteSpace also offers potential in identifying emerging science, with tools to detect and depict trends and "burstiness." It offers a rich repertoire of ready-made and adaptable procedures to tailor to one's particular needs, enabling data exchange to other software as well. This book offers meaningful and extensive help for a would-be user or an advanced user to go further in "cite space" -- along with nice case examples!"

Alan Porter, Georgia Institute of Technology, USA

"This manual is also a textbook about what is possible using scientometric data. Chaomei Chen continues to be deeply invested in making this new version of CiteSpace increasingly accessible to colleagues and students. The suite of analytical techniques is admirably comprehensive!"

Loet Leydesdorff, University of Amsterdam, The Netherlands

"Chaomei Chen has written an easy-to-read and insightful manual on science mapping based on his powerful Citespace software, useful for novices and experts alike."

Henry Small, SciTech Strategies Inc., USA

"The book is quite certainly invaluable for anyone interested in using CiteSpace – who better to give details of the various implemented techniques and algorithms than the author of the tool itself. This book will also be quite helpful in courses structured around the analysis and modeling of Complex Adaptive Systems such as found in the domain of scientific literature, paper authors, journals, and institutions."

Muaz A. Niazi, COMSATS Institute of IT, Pakistan

nova publishers
www.novapublishers.com

CiteSpace: A Practical Guide for Mapping Scientific Literature ♦ Chen

NOVA

Chaomei Chen

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

CiteSpace

A Practical Guide for Mapping Scientific Literature

NOVA

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

CITESPACE

A PRACTICAL GUIDE FOR MAPPING SCIENTIFIC LITERATURE

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

Additional books in this series can be found on Nova's website under the Series tab.

Additional e-books in this series can be found on Nova's website under the eBooks tab.

COMPUTER SCIENCE, TECHNOLOGY AND APPLICATIONS

CITESPACE

A PRACTICAL GUIDE FOR MAPPING SCIENTIFIC LITERATURE

CHAOMEI CHEN

New York

Library of Congress Cataloging-in-Publication Data

ISBN: 978-1-53610-280-2

Published by Nova Science Publishers, Inc. † New York

To Baohuan, Calvin, and Steven

CONTENTS

PREFACE

The idea of being able to interact with visually represented patterns and trends of how scientific knowledge has come along is intriguing and fascinating. Do we even remember how many times we wish that a comprehensive and up-to-date systematic review of the topic is out there when we need it? How many times do we realize that after turning over the last page of a promising review, the topic of our interest is either missed altogether or only covered superficially.

Fundamentally, using CiteSpace is transforming not only the way you learn about a scientific domain but also the way you think as a problem solver, a discoverer, and/or a communicator. Now you have more alternative ways to learn about the state of the art of a field, a discipline, or several intermingled disciplines. Your access to the rich body of scientific literature is no longer limited to systematic reviews written by well established but always busy gurus. You free yourselves from chasing up evasive details in hundreds of, or more likely, thousands of seemingly relevant papers either freshly printed or quietly hidden in a blind spot of our otherwise comprehensive literature search.

CiteSpace was initially released as a research prototype on Sept 25, 2003. Having it wrapped up and tucked in under a graphical hood has turned out to be a good decision in several ways. The system has been instrumental to serve an organizational role in accommodating various pieces of software and making them to do something interesting and useful. About 20 years ago, we published our first meta-analysis of hypermedia. We certainly didn't know anything about Heilmeier's questions back then. In fact, I am not even sure whether Heilmeier knew about his own questions at that point. The meta-analytic thinking was intriguing. If numerous research studies investigated the

same issue, they would settle on some consensus. You would think, wouldn't you? The diversity is found in theoretical foundations, methodologies, analytic reasoning and interpretation, considering they are supposed to address the same issue. So is ambiguity and uncertainty. We published our second meta-analysis in 2000, this time on information visualization. I came to realize that keeping abreast of the state of the art of a scientific domain must deal with the endless assessment-action-reassessment pattern. For better or worse, the threshold of getting published has been lowered steadily. The rate of retraction, on the other hand, is increasing. Unless we can shift our burden to computational tools, both of the quality and productivity of the research community as a whole would be jeopardized given the alarmingly fast growing volume of the potentially relevant information that we ought to check it out.

CiteSpace has several remarkable moments. For instance, with patterns and trends visualized by CiteSpace, we detected a shift of focus at the disciplinary level in research on mass extinctions. Through distinct patterns, we learned how various fields are connected in the broad context of terrorism research. More joyfully, the research that topped our chart was awarded the Nobel Prize in Medicine five months after the publication of our study. I have also received exciting emails from people I couldn't even pronounce their names telling me how glad they were seeing where a CiteSpace-guided tour led them to find. CiteSpace can get personal too. One of the retirement gifts to a research institution's director was the trajectories of his research and how they impact today.

CiteSpace has been evolving over more than a decade. Its name reminds the same, but many of its components have been upgraded and many new components are being added. Writing a comprehensive manual has been on my agenda for too long. In part, the system transforms itself faster than my writing. With more users attracted to the community of mapping scientific frontiers, it is a good time to provide a self-contained and almost jargon free book to explain the abstract and visionary paradigm of understanding scientific knowledge as it moves along.

It is worth noting that this is a particularly interesting time for anyone who is interested in mapping scientific literature. New tools frequently emerge. The diversity is widening the spectrum of our options as well as our horizons. After all, open mindedness breads creativity.

CiteSpace 2.0 was released in 2005, followed by 3.0 in 2011 and 4.0 in 2015. And now 5.0 in 2016. According to CiteSpace's What's New, this is the 350^{th} documented update. I hope this book will make it easier for you to

navigate through the intellectual space chartered by CiteSpace. As always, if you have any questions or need any help, you can find me in several cyberlocations (see Chapter 1 for detail).

I'd like to take this opportunity to say thank-you to a few people to whom I am particularly grateful. Pak Chong Wong, Pacific Northwest National Lab, for your support as an editor, a co-author, a co-chair, and a collaborator. Rod Miller, Keene State College, for many in-depth discussions on research, strategic planning, and vision. Mike Taylor, Digital Science, for thoughtful conversations on research topics that I am always fascinated with. Jie Li, Shanghai Maritime University, for testing various versions of CiteSpace and for organizing training workshops.

A big thank-you to my wife Baohuan for everything. Special thanks to Steven Chen for proofreading.

Chaomei Chen
Drexel University
8/31/2016

Chapter 1

INTRODUCTION

The constantly growing body of scholarly knowledge of science, technology, and humanities is an asset to mankind. Although new discoveries expand the existing knowledge, they may concurrently render some of it obsolete. It is crucial for scientists and other stakeholders to keep their knowledge up to date.

This chapter begins with Heilmeier's Catchism to set the stage for critical questions scholars and researchers need to answer and how CiteSpace can help.

1.1. HEILMEIER'S CATECHISM

Funding agencies such as DARPA (Defense Advanced Research Projects Agency) encourage researchers to formulate and communicate their research ideas through a list of questions, collectively known as Heilmeier's Catechism. The most common version of the list is as follows:

1. What are you trying to do? Articulate what you want to achieve using no jargon.
2. How is it done today, and what are the problems?
3. What is new in your approach and why do you think it will do better?
4. Who cares?
5. If you are successful, what difference will it make?
6. What are the risks and the payoffs?
7. How much will it cost?
8. How long will it take?

9. What are the midterm and final "exams" to check for success?

So what are we trying to do? Our goal is to make it easier and more efficient to answer the second question: how is it done today? In fact, our goal is to answer not only how it is done today but also how it has been done so far. The "it," an intentionally flexible entity, will be referred to as a knowledge domain here, and can pertain to a research topic, scientific field, or multiple disciplines. Later in the book, we will discuss further about how to do it by collecting the right data.

Traditionally, researchers rely on their own reading and systematic reviews published in the literature to obtain a good understanding of a subject. A good systematic review typically provides an overview of the subject matter's history, major research questions, landmark studies, established methods and techniques, and remaining challenges. Complete reliance on systematic reviews, however, is not always realistic or feasible. For example, systematic reviews may not even exist for an emerging field of study. Existing systematic reviews may be outdated and may not match our interests as closely as desired. While waiting for the next systematic review to meet our needs, wouldn't it be nice to have an alternative way to keep abreast of the scientific literature as the knowledge domain continuously matches on?

CiteSpace is designed to provide such an alternative so that we can use our own datasets to answer questions about an ever-changing knowledge domain (Chen 2004; C. Chen, 2006). As a computational tool, we will be able to use it as needed. Traditional systematic reviews are typically written by a small number of domain experts, who have their own specialized areas of interest. They have their own favorite perspectives of the world, which may or may not be compatible with those of other domain experts in the field. CiteSpace reduces such biases considerably by embracing the publications of authors across a wide spectrum of perspectives, schools of thought, and disciplines. There is no restriction in terms of the discipline of such publications. In fact, CiteSpace has been applied to the studies of over 60 different scientific fields.

The unit of analysis in CiteSpace is a knowledge domain. Thomas Kuhn's structure of scientific revolutions fundamentally influenced CiteSpace's design. According to Kuhn, sciences advance through various stages, such as normal science, crises, and paradigm shifts (Kuhn, 1962). In the normal science stage, the research agenda is clear and the foundation of the current thinking is considered reasonably sound. Crises arise when such foundations become questionable, leading to the appearance of competing paradigms.

Previously dominating paradigms may lose their positions to competing paradigms in what are known as paradigm shifts.

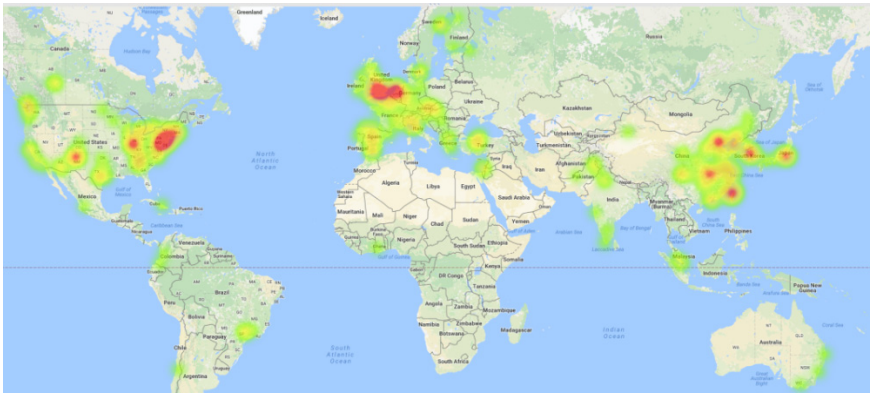

Figure 1. CiteSpace has been used worldwide.

Kuhn's theory provides a fitting framework to derive our strategies for conducting our own systematic review. We will be looking for trajectories of competing paradigms, turning points of paradigm shifts, long-lasting research interests, and transient bursts of scholarly activities.

Ben Shneiderman has a well-known mantra that characterizes the process of visual information retrieval: overview first, zoom and filter, then details-on-demand (Shneiderman, 1996). To a great extent, Shneiderman's mantra is applicable to a conceptual journey that aims to charter the landscape of a scientific domain's literature. To understand the status of a scientific domain, one must grasp an overview of the domain first. What are the major components of the domain? How are they connected with each other? Once we have the big picture of the domain, the next step is to select patterns that would be particularly worth exploring. Which component attracts the current attention of the scientific community? Which one is diminishing? We may be interested in reasons why a paper becomes a landmark.

We need a variety of information to guide ourselves as we navigate through the landscape or the space of knowledge. The types of information we need and how well we can make use of them depend on many factors. A visitor traveling in a city for the first time relies on landmarks much more than a visitor who has been visiting the city several times. It is less likely for a local resident to resort to a landmark to navigate, and an adventurous explorer may deliberately choose the most unusual paths to travel. Our prior knowledge and

experience plays a similar role in navigating through the invisible space of scientific knowledge. Although domain experts and novices will probably focus on different types of patterns, we have great freedom in choosing the course of our journey in the collectively constructed intellectual space.

1.2. HOW CAN CITESPACE HELP?

The design of CiteSpace is tailored to answer questions about the structure and dynamics of a knowledge domain. Here are some typical questions:

1. What are the major areas of research based on the input dataset?
2. How are these major areas connected, i.e., through which specific articles?
3. Where are the most active areas?
4. What is each major area about? Which/where are the key papers for a given area?
5. Are there critical transitions in the history of the development of the field? Where are the 'turning points'?

As mentioned earlier, CiteSpace's design is inspired by Thomas Kuhn's structure of scientific revolutions. The central idea is that focal points of research change over time, either incrementally or significantly. By studying the footprints uncovered by scholarly publications, we can trace science's development over time.

Members of the contemporary scientific community make their contributions. Their contributions form a dynamic and self-organizing system of knowledge. The system contains consensuses, disputes, uncertainties, hypotheses, mysteries, unsolved problems, and unanswered questions. Studying a single school of thought is insufficient. In fact, a thorough understanding of a specific topic often relies on understanding its relations to other topics.

The foundation of the CiteSpace is network analysis and visualization. Through network modeling and visualization, you can explore the intellectual landscape of a knowledge domain, discern what questions researchers have been trying to answer, and discover the methods and tools they have developed to reach their goals.

Similarly, and probably to a greater extent, CiteSpace can generate X-ray photos of a knowledge domain, but to interpret what these X-ray photos mean, you need to have some knowledge of various elements involved.

This is not a simple task; rather, it is conceptually demanding and complex. If you are about to write a novel, the word processor can make writing and editing easier, but it can hardly help you to create the plot or enrich the character of your hero. Similarly, CiteSpace can generate X-rays of a knowledge domain, but it cannot interpret the X-rays.

The role of CiteSpace is to shift some of the traditionally labor-some burdens to computer algorithms and interactive visualizations so that you can concentrate on what human users are best at in problem solving and truth finding. However, it is probably easier to generate some mysterious looking visualizations with CiteSpace than to fully understand what those visualizations depict.

1.3. GEOGRAPHIC DISTRIBUTION OF USERS

CiteSpace has been used by people at many cities around the globe. The map below shows these cities. The size of a circle represents the number of distinct IP addresses of CiteSpace users at a city. The use of CiteSpace here means a user actually interacted with CiteSpace. The color of a city indicates the average number of days from the first use to the most recent use. The longest average duration is shown in red. Shorter ones are shown in blue.

Beijing has the largest number of unique IPs and the longest average duration of use. The highest concentrations of CiteSpace users come from China, European countries, the US, and Brazil.

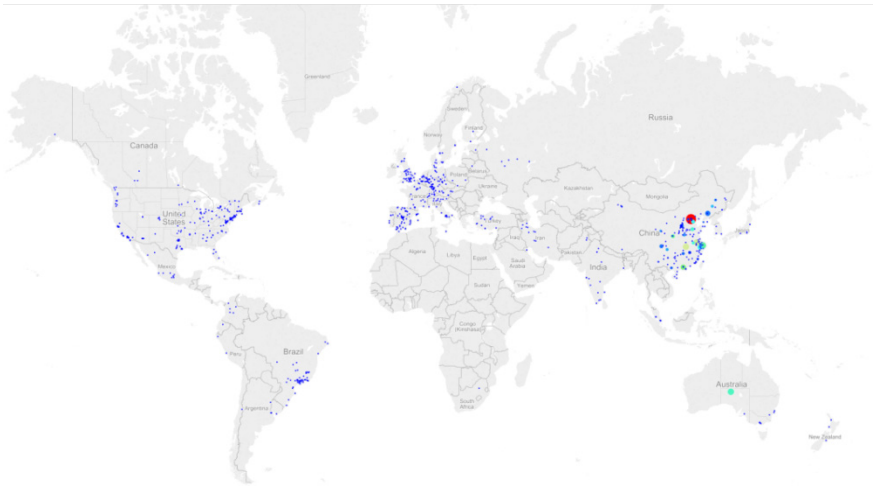

Figure 2. The geographic distribution of CiteSpace users.

1.4. How Should I Cite CiteSpace?

There are three representative technical publications of CiteSpace and many case studies and applications of CiteSpace to specific domains. The three key publications are the 2004 PNAS paper on the initial design of CiteSpace, the 2006 JASIST paper on CiteSpace II, and the 2010 JASIST paper on CiteSpace III.

The 2004 PNAS paper is the initial publication on CiteSpace (Chen 2004). The 19-page 2006 JASIST paper gives the most thorough description of CiteSpace II's key functions and two case studies (C. Chen, 2006), plus a follow-up study of domain experts identified in the visualizations. The 24-page 2010 JASIST paper describes technical details on selecting cluster labels and compare the results with labels chosen by domain experts (C. Chen, F. Ibekwe-SanJuan, & J. Hou, 2010).

In addition to the three key publications, dual-map overlays are introduced in a 2014 article (Chaomei Chen & Leydesdorff, 2014), Structural Variation Theory (SVT) in (C. Chen, 2012), and a study of retraction in a 2013 article (C. Chen, Hu, Milbank, & Schultz, 2013). We have applied CiteSpace in the study of a diverse range of domains such as regenerative medicine, orphan drugs, mass extinctions, terrorism research, human computer interaction, string theory, and astronomy. Other users have applied CiteSpace to numerous domains.

Table 1. Three key technical publications of CiteSpace

Google Scholar Citations	Reference	Highlights
1335	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.	CiteSpace II
582	Chen, C. (2004) Searching for intellectual turning points: Progressive Knowledge Domain Visualization. *Proc. Natl. Acad. Sci. USA* 101(Suppl.): 5303-5310.	CiteSpace I
276	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.	CiteSpace III

Table 2. Additional functionalities and applications of CiteSpace

Citations	Reference	Highlights
82	Chen, C., Hu, Z., Liu, S., Tseng, H. (2012) Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace. *Expert Opinion on Biological Therapy*, 12 (5), 593-608.	Application
48	Chen, C. (2012) Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, 63(3), 431-449.	Structural Variation Analysis (SVA)
26	Chen, C., Dubin, R., Kim, M. C. (2014) Emerging Trends and New Developments in Regenerative Medicine: A Scientometric Update (2000-2014). *Expert Opinion On Biological Therapy,* 14 (9), 1295-1317.	Application
25	Chaomei Chen, Zhigang Hu, Jared Milbank, Timothy Schultz (2013) A visual analytic study of retracted articles in scientific literature. Journal of the American Society for Information Science and Technology, 64(2), 234-253.	Study of retraction
20	Chen, C., Leydesdorff, L. (2014) Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. *Journal of the American Society for Information Science and Technology*, 65(2), 334-351.	Dual-Map Overlays

Table 3. Resources on CiteSpace

Resource Name	URL	Type
CiteSpace Home	http://cluster.cis.drexel.edu/~cchen/citespace/	Website
How to Use CiteSpace	https://leanpub.com/howtousecitespace	e-Book
CiteSpace101	https://sites.google.com/site/citespace101/	Website
Facebook	https://www.facebook.com/CiteSpace-276625072366558/	
Twitter	http://twitter.com/CiteSpace	
Blog (in Chinese)	http://blog.sciencenet.cn/home.php?mod=space&uid=496649	
Collect Data	https://youtu.be/D75An6DqsfA	Video
Get Started	https://youtu.be/zL2oFRqiP6k	Video
Dual-Map Overlays	https://youtu.be/4D6gxn_l6k8	Video
Tutorial in Portuguese	https://youtu.be/Gqx0fvSSq9Q	Video
Google+	https://plus.google.com/116966090977405653251/posts	Photos
ResearchGate	http://www.researchgate.net/profile/Chaomei_Chen/	Papers

In addition to this book, there are several online resources where you can find various relevant information and post your questions.

CiteSpace is designed to make it easy for you to answer questions about the structure and dynamics of a knowledge domain. Here are some typical questions:

- What are the major areas of research based on the input dataset?
- How are these major areas connected, i.e., through which specific articles?
- Where are the most active areas?
- What is each major area about? Which/where are the key papers for a given area?
- Are there critical transitions in the history of the development of the field? Where are the ‘turning points’?

The design of CiteSpace is inspired by Thomas Kuhn’s structure of scientific revolutions. The central idea is that centers of research focus change over time, sometime incrementally and other times drastically. We can trace the development of science by studying the footprints uncovered by scholarly publications.

Members of the contemporary scientific community make contributions form a dynamic and self-organizing system of knowledge. The system

contains consensuses, disputes, uncertainties, hypotheses, mysteries, unsolved problems, and unanswered questions. It is not enough to study a single school of thought. In fact, a better understanding of a specific topic often relies on an understanding of how it is related to other topics.

Chapter 2

BASIC CONCEPTS AND PRINCIPLES

The design and application of CiteSpace involves fundamental concepts and principles from several fields of study. For example, citation indexing is originated in bibliometrics and scientometrics. The differences between the h-index and the g-index have implications on our design decisions. Properties of networks and their practical implications are also introduced in this chapter. A diverse range of theories have influenced the design of CiteSpace. A good understanding of these concepts, theories, and their practical implications explained in this chapter is essential for a rewarding experience with CiteSpace.

2.1. CITATIONS

In the scholarly world, a citation refers to the fact that an article A makes reference to an existing article B. A is also known as the source of the citation, or simply the citing article. B is called a reference or the target of the citation. The link established by the A→B citation is called a referential link. The citations of an article refer to the number of articles that cite the article. Usually, what is implicit but clear from the context is that the citations of an article are functions of the time elapsed since its publication and the scope of the sample articles under consideration. For instance, an article on Google Scholar is likely to have about 2 or 3 times more citations than its counterpart in the Web of Science.

What does the citation count of an article really represent? The consensus is that the higher the citation count for an article, the more likely the article is

valuable (Garfield, 1955). Of course, a less conservative but more controversial definition claims that the higher the number of citations, the higher the impact. Then this definition leads to a deeper question: what would qualify as an impact? What is the best way to measure and quantify an impact? Is it adequate to consider the impact of a publication alone? What about the impact of making a computational tool such as Pajek (Batagelj & Mrvar, 1998) available for free? Making CiteSpace a freely available tool for people who may otherwise have no first-hand access to visualizing and analyzing salient patterns in scientific literature is inspired by the generosity of pioneers who have made their tools available to others. Remarkably, Pajek is on the top of the list of such tools. While tools that appear later on, for instance, Gephi, become increasingly more powerful and aesthetically more pleasing, what pioneers chose to do has a profound and long-lasting impact.

Expectations on the level of citations differ from disciplines as well as fields, specialties, and groupings at finer levels of granularity. An article in biomedicine or physics may attract 1,000 citations within a year, whereas an article in an area with a much smaller community of researchers may have citations growing at a much slower rate. Does it make any sense to compare a swimmer's 1500m and a runner's 1500m? Field normalization has been proposed by many scientometricans to make citations more comparable across different fields. Whether one should compare citations across the fields in the first place is another subject.

From the baseline of article-level citations, one can expand the idea in two opposite directions. Going upwards, we can talk about author-level, institutional-level, or journal-level citations if we group articles along with their citations by author, institution, or journal. Going downwards, we can differentiate citations made in specific sections of an article and citations made to serve specific rhetorical or argumentative functions.

One may choose to split citations between multiple co-authors if attributing everything to the first author bothers us that much. In recent years, coauthorship has become increasingly sophisticated, especially with the conception and adoption of new species such as co-first authors. Should the first co-first author get more credit than the second co-first author? One thing for sure at least in this book, this is beyond the scope of our interest.

In surveying the knowledge space of a scientific domain, information such as a citation count directs our attention to potentially important areas. In terms of Ben Shneiderman's mantra for visual information retrieval, such information allows us to not only select the target for zooming and filtering in

a given forest, but also select the trees that shape up the forest in the first place.

2.2. CITATION INDEXING

Scientific literature consists of two types of articles – classic and transient. Classics are well-established and well-known contributions. Their positions in scientific literature are relatively stable, and enjoy high visibility. Success breeds success, as the Matthew Effect would predict. In contrast, the transient part of the world is volatile – things move fast and many articles are forgotten immediately after publication. Occasionally, some of these vanishing articles can be re-discovered, years after their initial publication. Such lucky ones are affectionately named « sleeping beauties ». It is particularly hard for computational models to predict where the next wake-up kiss might come from.

A citation index is the idea to trace citations made by authors in their articles in order to identify connections that could be otherwise missed – for example, by searching for matching vocabularies alone. Eugene Garfield, the pioneer of citation indexing, repeatedly emphasizes that his original intention to create scientific citation index was to help people find more relevant articles in the literature (Garfield, 1955). Using citations (especially journal-level citation metrics) as the basis of research evaluation was never his idea of the silver bullet that many of us today are eagerly and blindly embracing.

2.3. MEASURING THE QUALITY AND PRODUCTIVITY

In addition to raw counts of citations, numerous indicators have been proposed and continuously tweaked in attempts to signal something worthy of a second look and sustaining an important decision. The h-index is clearly very impactful one to date (Hirsch, 2005, 2007). The h-index is a number that can be derived from the research portfolio of an individual, institution, country, or arbitrary set of publications. The magic number h is the number of publications in the sample that have been cited h or more times.

Consider an experienced and active researcher E who has published a lot of papers. Some of E's papers have been cited a lot, while other papers have been cited less frequently or not cited at all. Researcher N is a newbie to the same domain in which E has been active for decades. N has a handful of

publications, and most of them appeared in the last 6 months. The h-index of E would be much higher than the h-index of N. Thus, the h-index signifies the superiority of E over N in terms of productivity and impact.

Criticism of the h-index is probably as plentiful as praise. New indices have spun off from the original h-index so fast that one begins to worry how long a new index can still be named from the English alphabet. We will just introduce one refinement of the h-index – the g-index (Egghe, 2006).

CiteSpace allows users to select nodes based on their g-index values, but not h-index values. The primary difference between the h and the g indices is that the g-index accounts for citations of the articles in the sample, whereas the h-index doesn't. In addition to raw counts of citations, the g-index provides an alternative way to select references and other types of entities to be featured in the subsequent visual analytic process.

2.4. Representing a Knowledge Domain

In CiteSpace, a knowledge domain is modeled by various types of time series of networks. Imagine that CiteSpace is an X-ray machine. Instead of taking an X-ray of the chest, we are interested in taking an X-ray of the scientific literature. Since the underlying knowledge domain is changing all the time, it is unlikely that one snapshot will give us enough information to learn about the domain. So we take a series of snapshots across a period of time that is long enough to reveal interesting patterns such as those revolutionary turning points or the norm of a normal science. Typically, we take one snapshot per year. Luckily, we don't have to wait for another year to take another snapshot. The information is in our data. CiteSpace can sort out the data so that a series of snapshots can be taken consecutively in a single process. CiteSpace will then visualize these snapshots in a way that allows us to find answer to questions concerning a knowledge domain in the broadly defined framework of Kuhnian paradigms.

2.4.1. Connectivity

A scholarly publication typically includes a list of references. These references are mentioned in the body of the article as part of the narrative. References on the same list are co-cited by the host article, i.e., the citing article, or simply the citer. Finer-grained co-citations are possible if more data

such as the full text of the article is available. For instance, two references can be co-cited at the sentence, paragraph, or article level. Without losing much of generality, we refer to the article-level co-citations unless we explicitly mention other levels of granularity.

The basic idea is that if two items are frequently co-cited or co-occurring, then there is a good chance they are related. This is indeed the basis of co-citation analysis. Co-citation relations are local events because one co-citation instance involves only two references or two entities. The power of co-citation analysis is that these local relations can be aggregated to reveal global patterns such as a long path across disciplinary boundaries or a clusters of references that represent the footprint of a Kuhnian paradigm.

Various visual encodings are designed in CiteSpace to direct attention to the most important signals so that we can capture the most important patterns easily and effectively. Each snapshot made by CiteSpace is a network. A network is a set of interconnected entities. The entities are known as nodes or vertices. The connections are known as links or edges, and each connection can be directed or undirected. For example, two researchers who published an article together are co-authors. Their co-authorship is an undirected connection. In contrast, an article citing a reference makes a directed referential connection between the article and the reference. The strength of a connection can be modeled either simply by a status of yes or no or by a real number, typically normalized over the unit interval [0, 1]. The value that represents the strength of the linkage is also known as the weight of the link, with 1.0 being the strongest and 0.0 indicating no connection.

Each network has the largest connected component. It is the largest sub-network in which you can start from any node and reach any other node. A clustering algorithm may still find that nodes of the largest connected component belong to more than one cluster. It is a common practice in network analysis to focus on the largest connected component.

2.4.2. Structural Holes and Good Ideas

The structural hole theory was introduced by Ronald S. Burt of the University of Chicago (Burt, 1992, 2004). The problem studied by Burt is whether there is a meaningful connection between people's ideas and their positions in a social or professional network. He found evidence of connections based on a concept called structural holes. In a fully connected social network, everyone is connected to everyone else. Therefore, information

flows freely from anyone to anyone else. In such networks, there is no structural hole.

In contrast, social networks have structural holes if a free flow of information is impossible. In other words, individuals of the social network are not equal in terms of how much information they may receive because how they are connected to others in the network. Missing links that prevent them from receiving as much information as others form structural holes. Burt found out that people around a structural hole tend to have competitive advantages over people positioned elsewhere on the network. The whole idea boils down to whether one is exposed to a diverse spectrum of information, opinions, or views.

The role of a structural hole in a social network is extensible to other types of networks where the flow of information can be affected by a network's connectivity pattern. In the context of studying scientific literature, the idea of structural holes can be translated into an important property of a node in a network – its betweenness centrality. A node with a strong betweenness centrality score has a great influence on how information flows through the node.

There are several types of centrality measures for nodes in a network, including betweenness centrality and degree centrality (Brandes, 2001; Freeman, 1977). The betweenness centrality of a node measures the extent to which paths in the network may go through the node. This idea is similar to measuring the importance of a tollbooth located in the traffic network. A tollbooth on highway I-95 between Philadelphia and New York would be more important than a tollbooth on a highway with a much lighter traffic, for example, I-476. Furthermore, Philadelphia and New York both function as hubs linking to other big cities. As a result, the betweenness centrality of a tollbooth on I-95 should be relatively high in the network of highway traffic. A node with a high betweenness centrality would be particularly informative in understanding why two clusters are connected.

2.4.3. Information Foraging

The betweenness centrality of a node helps us to identify structural holes where good ideas are more likely to appear than other areas of the network. In reality, merely having good ideas is probably not good enough. People need to make their own decisions and make up their minds. An inspiring theory is called optimal information foraging theory (Peter Pirolli, 2007; P. Pirolli &

Card, 1999). The theory was originally proposed by Pete Pirolli to explain the decision making process when people search for information. The optimal information foraging theory itself is inspired by an optimal foraging theory. When we search for information, we need to make a series of decisions. All these decisions serve a simple purpose: we need to get things done as efficient as possible. Pirolli uses the notion of profitability to quantify the efficiency, which is essentially a gain-to-risk ratio. According to the theory, all our decisions in a foraging process try to maximize the ratio of the expected gain to the potential risk. Sometimes it is hard to justify taking a high risk unless the expected gain is high enough. New observations may reduce the level of a perceived risk if others have been successful on the same or similar tasks. In a scientific field, the publication of a high-risk idea usually is followed by more studies. It is therefore very likely that the initial publication has reduced the perceived risk and thus increased the overall profitability of doing the same thing or something similar. As a result, it seems to be reasonable to impose two conditions on a potentially valuable idea: it needs to be around structural holes and it should have followers. We can deal with the former with betweenness centrality scores and the latter with the citation burstness or any other types of burstness.

2.4.4. The Burstness

Burstness measures the rate of change (Kleinberg, 2002). The burstness of the frequency of an entity over time indicates a specific duration in which an abrupt change of the frequency takes place. In CiteSpace, citation burst and occurrence burst are both supported.

In CiteSpace, you can find out the burst of citations to a reference. A node with a strong burstness usually indicates a potentially interesting work that has attracted significant attention within a short period of time.

The role of CiteSpace is to shift some of the traditionally laborious burdens to computer algorithms and interactive visualizations so that you can concentrate on what human users are best at in problem solving and truth finding. However, generating some mysterious visualizations with CiteSpace is easier than fully understanding what those visualizations reveal and who may benefit from such findings.

A composite metric sigma is defined in CiteSpace to measure the combined strength of structural and temporal properties of a node, namely, its betweenness centrality and citation burst (C. Chen et al., 2009). A node with a

high sigma value has not only a strategically important structural property but also special temporal implication.

2.4.5. Modularity

CiteSpace stitches snapshots taken at adjacent time points to form a merged network, i.e., a panorama of the domain in question. A visual analytic process with CiteSpace primarily focuses on patterns revealed by the merged network. A network represents how a set of nodes are interconnected. Individual nodes are interesting – each node may have several attributes and each attribute may lead to insightful details. If we focus too much on individual nodes and other local details, however, we may lose sight of the forest for the trees. An important step in the visual analytic process of CiteSpace is to divide the panoramic network into groups of individual nodes. Given a network, this kind of division can be done with a diverse range of clustering algorithms. The groups identified are called clusters. The most commonly used criterion is that nodes within the same cluster should be much more tightly coupled than nodes from different clusters. Some clustering algorithms allow overlapping clusters, which means one node may simultaneously belong to multiple clusters, whereas other algorithms are rather strict – each node belongs to one and only one cluster. Clusters in CiteSpace follow the latter – no overlaps!

The modularity of a network measures the extent to which a network can be decomposed to multiple components, or modules. This metric provides a reference of the overall clarity of a given decomposition of the network.

If a network's modularity is close to 1.00, then the network is clearly divided into distinct groups. In contrast, if its modularity is below 0.30, one would expect to see many between-cluster links. We should be careful in interpreting these between-cluster links. Between-cluster links imply that the clustering algorithm in our hand is unable to separate nodes in the network thoroughly. However, the inability to separate the network's nodes may reflect the hidden structure of the knowledge domain, provided the data sample is representative enough.

A network's modularity is a global measure of the overall structure of the network (Newman, 2006; Takeda & Kajikawa, 2010). As the underlying structure changes, the modularity of the network is likely to change. Two types of changes are of particular interest: changes that essentially involve a single cluster and changes that involve multiple clusters. The latter is likely to induce a much more dramatic change in terms of the network's modularity than the

former. In other words, changes that involve multiple clusters are likely to have a more profound impact than those limited to individual clusters. We can detect such changes by monitoring how modularity values change in response to new information received by the system.

Let's pursue this line of reasoning further. Consider the body of domain knowledge accessible to its associated community of researchers, ranging from pioneers several generations ago to the current generation. The body of knowledge is an adaptive complex system. It is adaptive because new discoveries and new ideas may change our beliefs and behavior, and is complex because its input and output are not linearly related. On the one hand, a large portion of the scientific literature is transient, meaning this part of the literature essentially attracts no attention from the scientific community. On the other hand, a very small number of publications can attract the attention of the scientific community quickly and dramatically. The million-dollar question is to what extent, if possible at all, one can identify any predictable characteristics of the very small group of publications.

If the publication of a new article serves as a new signal sent to the adaptive complex system, then the system may either remain intact or respond drastically. If we measure the modularity of the system, we should expect to see either a dramatic change in modularity or little to no change.

Focusing on the change of modularity induced by a newly published article is the basis of our Structural Variation Theory (SVT) (C. Chen, 2011). CiteSpace provides a function to track how modularity values change year by year as the publication of articles in the sample dataset unfolds.

The silhouette value of a cluster measures the quality of a clustering configuration. Its value ranges between -1 and 1, with 1 representing a perfect solution. However, to ensure a sound interpretation in CiteSpace, both the modularity and silhouette scores should be taken into account when interpreting the results (C. M. Chen, F. Ibekwe-SanJuan, & J. H. Hou, 2010).

A cluster with a perfect Silhouette score can be deceptive if the formation of the cluster is the artifact of citations made by a very small number of articles. It is therefore a good idea to double check how many citing articles are associated with a particular cluster before attempting to interpret the nature of a cluster simply based on a strong silhouette score.

Chapter 3

GETTING STARTED WITH CITESPACE

This chapter introduces the initial configuration of the software and several types of visual analytic procedures supported by CiteSpace. Practical details of how to manage a CiteSpace project and how to optimize a CiteSpace session are also explained. The chapter also outlines the overarching strategy that consists of multiple analytic processes, ranging from geographical to thematic and from structural patterns to temporal patterns. Ultimately, our study is driven by Heilmeier's questions.

3.1. DOWNLOAD

CiteSpace is a Java application. In theory, with a Java application, you just need to write it once, and run it anywhere. In reality, your computer must have Java installed. The Java Runtime Environment (JRE) should be sufficient, while Java Development Kit (JDK) will be overkill for just running CiteSpace. Java applications are bytecodes that need to be translated for your computer to execute. In order to run CiteSpace, we need to run a Java Virtual Machine (JVM) and the JVM will then get CiteSpace running.

If your computer does not have Java installed, you need to download and install Java first. You will also need to know whether you have a 32- or 64-bit operating system in order to download the right version.

CiteSpace is downloadable as a single file in the 7z format from the CiteSpace download page[1]. This is a compressed file. Uncompress the file to

[1] http://cluster.ischool.drexel.edu/~cchen/citespace/download/.

your computer. You will see a jar file, which is the CiteSpace proper, and a few batch files that you can use to launch CiteSpace. You will also see a StartCiteSpaceLargeChinese.bat file, which presets the Chinese locale especially for handling Chinese data files or if your file paths contain Chinese characters.

The files with .bat as their file extension names are batch files. Their role is to configure the JVM and start CiteSpace. You can edit these batch files with a text editor such as Notepad++. The batch files are very similar in content, but differ in how they configure the amount of RAM for the JVM.

Each version of CiteSpace has an expiration date to encourage you to use the latest version, which usually provides the most comprehensive set of functions and the least number of bugs. Some special edition versions, such as 5.0.R1 SE and 4.0.R5 SE, will remain valid for two or three years.

```
@ECHO OFF
ECHO ************************************************************************
ECHO *                                                                      *
ECHO *                   CiteSpace                                          *
ECHO *                                                                      *
ECHO ************************************************************************
ECHO
ECHO You may optimize the performance of CiteSpace by adjusting the following JVM parameters.
ECHO In general, the more RAM, the better.
ECHO -Xms1g: request at least 1GB ram for Java Virtual Machine
ECHO -Xmx1g: request at most 1GB, depending on your computer
ECHO -Xss5m: request 5MB ram for Java stack
ECHO -Duser.country=US -Duser.language=en
ECHO -Duser.country=CN -Duser.language=zh
@ECHO ON
java -Duser.country=US -Duser.language=en -Dfile.encoding=UTF-8 -Xms1g -Xmx12g -Xss5m -jar CiteSpaceV.jar
```

Figure 3. The content of StartCiteSpace.bat.

In a batch file, lines begin with “ECHO” or “@ECHO” are comment lines, which have no effect on how you run CiteSpace. Their purpose is to provide you with some additional information on configuring your JVM, such as running on more RAM or using the UTF-8 encoding when reading from data files.

You can double-click on the batch file to start CiteSpace. At this stage, the most common problem is that your computer cannot allocate the amount of RAM to the JVM as requested in the batch file. You will have to edit the batch file before you try again.

3.2. CONFIGURING THE JVM

The amount of memory CiteSpace can use affects its performance. To analyze a large amount of records, therefore, you should request as much memory as possible for the JVM by modifying parameters in the batch file.

Open the batch file with a text editor and modify the –Xmx2g in line 16 of the file. -Xmx2g means that CiteSpace asks for a maximum of 2GB of RAM for the JVM. If you want it to be 10GB, change the term to –Xmx10g. Save the changes and restart the program. Brief notes about how to set RAM related parameters are in lines 10 and 11 of the batch file.

The default locale in the StartCiteSpace.bat is for English (United States), i.e., en_US:

- Duser.country=US
- Duser.language=en

Using CiteSpace with other locales may cause some problems, especially if a comma or a decimal point in your default locale has different meaning different from the English/US locale. Such discrepancies may cause links missing from the network. StartCiteSpaceLarge.bat sets the en/US as the default locale for CiteSpace.

You can use the –Duser.country and –Duser.language to set your default locale for the JVM. For example, use zh_CN for Chinese (China),

- Duser.country=CN
- Duser.language=zh

Or, use de_DE for German (Germany)

- Duser.country=DE
- Duser.language=de

You may notice that recent releases of CiteSpace are all labeled as 64-bit versions. If you have a 32-bit computer, you can still use the 64-bit CiteSpace. Pragmatically, the major difference between 32- and 64-bit versions of CiteSpace boils down to the architecture of your computer, whether it has a 32- or 64-bit operating system. On a 32-bit computer, you are limited to about 2GB of RAM, whereas on a 64-bit computer, the limit is much higher at

192GB. The theoretical upper limit is 17.2 billion GB. We don't need to worry about reaching this limit for a long time. The more RAM you can allocate for Java (and in turn for CiteSpace), the faster it will run, especially when you have a large set of data.

CiteSpace builds a network model from a raw dataset. The size of the network is typically smaller than the largest possible network that can be built from the data. In other words, CiteSpace typically processes a much larger set of records behind the scene than the resultant visualized network.

The total number of records in your downloaded dataset can be very large. CiteSpace processes each record in your data files. Some records will be filtered out using selection criteria such as a citation threshold. The more RAM you can make available for CiteSpace, the larger sized network you can visualize with a faster response rate.

Additional tasks may increase the processing time. Pathfinder network scaling, for instance, is computationally expensive – the algorithm slows down quickly as the size of the network increases. Bibliometricians in Granada, Spain published a faster algorithm based on a nice property of a Pathfinder network: a Pathfinder network is the set union of all the possible minimum spanning trees of the network. My advice is that you should apply Pathfinder network scaling on small-to-medium sized networks per slice, in the range of 50~500 nodes per slice. With a faster computer (or if you are willing to wait patiently), you can raise the number accordingly.

CiteSpace can automatically select labels for clusters. The process starts from the smallest cluster and walks its way up to the largest cluster. This process may take longer if you have many large clusters. If you look at the command line window, you will notice it slows down towards the end.

The completion time of cluster labeling correlates with the size of your dataset. For instance, if you are only considering the last 10 years of a dataset spanning 100 years, carving out the smallest possible dataset that covers the 10-year window will reduce the processing time considerably.

The largest network that CiteSpace can visualize depends on the hardware of your computer – specifically, the amount of memory accessible to the JVM. The number of nodes and the density of the network can influence the clarity of the resultant visualization.

Bear in mind that despite its large size, the most informative visualization is not necessarily from a network. The majority of the nodes will be barely noticeable; hardly more than a few pixels in the background; they. We should focus on articles that play critical roles in the domain.

The following visualization shows a co-citation network of 30,640 nodes and 133,975 edges. The network represents what Drexel University's publications cite. In fact, this is what the top 10 most cited Drexel papers cite between 2000 and 2014.

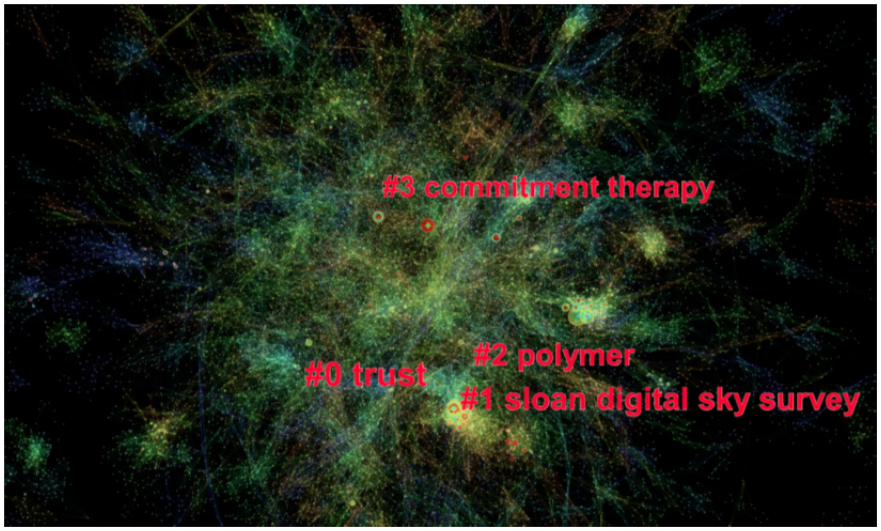

Figure 4. A visualization of what the top 10% most cited Drexel University's publications each year cite between 2000 and 2014. References=30,640, Instances of Co-Citation=133,975, Modularity Q=0.9756, Mean Silhouette=0.5702.

The galaxy-shaped concentrations are clusters of references that are often cited together. Red circles are references with citation bursts and indicate hotspots, which attract a substantial amount of attention within a short period of time. The labels in red are cluster labels, which summarize the essence of a cluster. For instance, the label *#1 sloan digital sky survey* is a clear giveaway – the cluster is formed by Drexel's researchers in astronomy and their collaborators.

3.3. LAUNCH

When you run CiteSpace for the first time, you may notice that CiteSpace is downloading additional files. These files include sample datasets for several Demo projects. As a beginner, you can use these Demo projects to learn basic functions of CiteSpace and various ways to configure a CiteSpace project. You

may of course choose to dive in directly to work on your own datasets. It can be rewarding as well as adventurous. Proper preparation can only help you to get the job done.

3.3.1. On Windows

On Windows, double click on the StartCiteSpaceLarge.bat file. A Windows command line (cmd) window will appear. If this is your first time running CiteSpace on your computer, CiteSpace will download some extra files, including some sample projects such as the Demo projects we will discuss later in the book. It usually should take less than a minute for CiteSpace to download all the files it needs. Sometimes, the server that holds these extra files may be temporarily unavailable due to maintenance, a power outage, or other reasons. If you see error messages about any missing files, you will have to let CiteSpace try again later.

The cmd window is a good source of information. If CiteSpace encounters a problem, it will send error messages to the cmd window. When you report a bug or a problem with your data, include the messages shown in the cmd window.

CiteSpace uses a special folder on your computer to store these extra files. On a Windows system, the folder is /Users/your_login_name/.citespace. If you want to refresh previous installations of CiteSpace, simply remove the .citespace folder and restart CiteSpace.

3.3.2. On Mac or Unix-based Systems

Since CiteSpace is a Java application, you can run it on a Mac or other computer systems as long as it has Java installed. The following example shows you the basic steps to get started with CiteSpace on a Mac.

Download CiteSpace to your Mac. Recent releases of CiteSpace are compressed in the .7z format. By default, Mac does not know how to handle .7z files. You will need to download software such as Unarchiver (which is free) to unpack the 7z file[2].

Once you unpack the 7z files, the simplest way to get started is to click on the CiteSpaceV.jar file while holding the Control key on Mac. Then select the

[2] http://osxdaily.com/2010/12/13/open-7z-files-on-a-mac/.

Open menu from the pop-up list. Due to Java security reasons, you will see a dialog box with two options: Open or Cancel. Choose Open to proceed. It will not harm your computer. After you choose Open, CiteSpace will begin to run on Mac. You will see its opening window. Press the **Agree** button to continue. It is a good idea to get familiar with the basic functions and the overall analytic process of CiteSpace before diving into the wonderland of your own data. In next chapter, we will introduce several Demo projects that come with the CiteSpace release package.

If you want to configure various Java Virtual Machine parameters in more detail, like what the batch files do on Windows systems, you should generate a bash file, which is the counterpart of a batch file on Mac. A bash file should have .sh as its file extension. It should be executable. Name the bash file as StartCiteSpace.sh, just to be consistent.

First, create a StartCiteSpace.sh file with the following two lines:

```
#!/bin/bash
java -Xms1g -Xmx4g -Xss5m -jar CiteSpaceV.jar
```

Then, make the .sh file executable by running the following command.

```
chmod +x StartCiteSpace.sh
```

To run the executable file, simply type its name at the prompt or double click on it.

3.4. SCIENCE MAPPING WITH CITESPACE

The diagram below illustrates the most essential visualization and analytic functions of CiteSpace. The primary source of input data for CiteSpace is the Web of Science. The orange route is the easiest and most intuitive one. You can generate collaborating author overlays on geographic maps and interact with the maps with Google Earth. With a few simple steps in Google Fusion Tables, you can also generate very cool heatmaps. The blue route leads to dual-map overlays, which show us the big picture of trajectories of a set of scholarly publications at the level of journals. The third type of overlays is called network overlays, which allow you to contrast multiple networks by superimposing one network over another. The purple route takes you to text analysis. Finally, the green route serves to fulfil the core function of CiteSpace

– conducting a new generation of document co-citation analysis – Progressive Knowledge Domain Visualization (pKDViz).

To launch CiteSpace, double click on the StartCiteSpace.bat or StartCiteSpaceLarge.bat file. A command prompt window will appear with various information on the status of CiteSpace and any errors it encounters.

You will then see an **About CiteSpace** window appear, displaying system information of your computer and the version of Java installed on your computer. To proceed, click on the Agree button, which gives your consent for CiteSpace to generate event logs for research purposes. After that, you will see the main user interface of CiteSpace.

The user interface is divided into left and right halves. The left half contains controls of projects (i.e., input datasets) and progress report windows. The right half side contains several panels for configuring the process with various parameters.

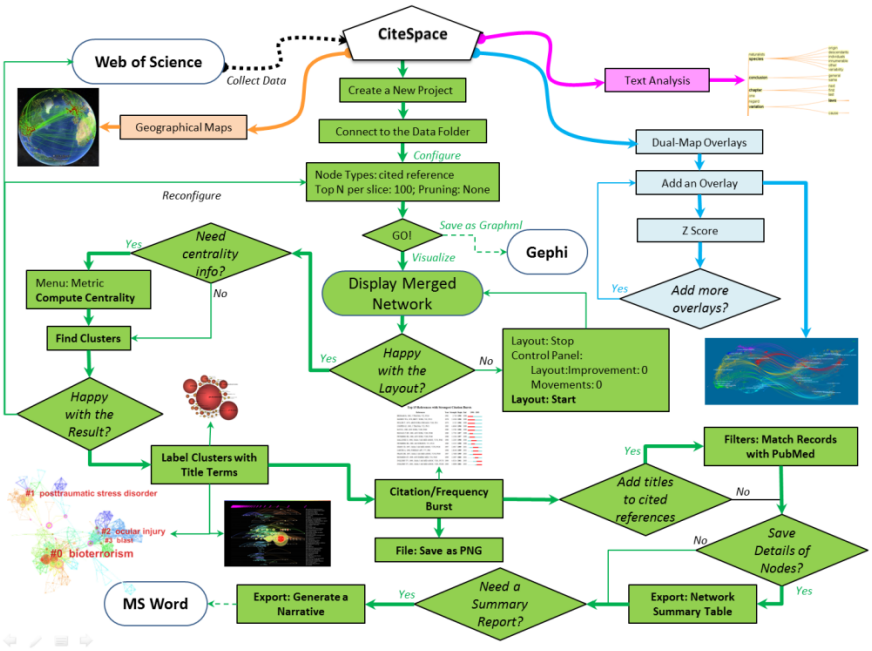

Figure 5. Major visual analytic paths supported by CiteSpace.

In a nutshell, the process in CiteSpace takes an input dataset specified in the current project, constructs network models of bibliographic entities, and

visualizes the networks for interactive exploration for trends and patterns identified from the dataset.

The demo project contains a dataset on publications about terrorism research. These bibliographic records were retrieved from the Web of Science. See later sections for tips on constructing your own dataset.

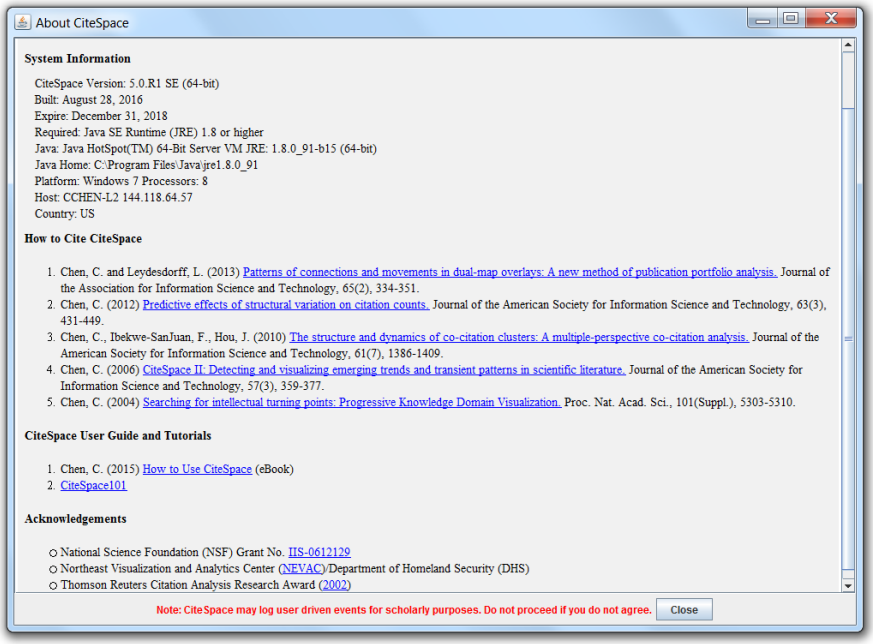

Figure 6. The About CiteSpace window.

You have seen some of the basic moves. CiteSpace has many other features. We will introduce other features at more advanced levels.

The diagram below illustrates the most essential visualization and analytic functions of CiteSpace. The primary source of input data for CiteSpace is the Web of Science. The green route is for document co-citation analysis and a series of other analyses. The blue route is for generating dual-map overlays. The purple route is for text analysis. The orange route is for generating geographic maps that can be viewed in Google Earth.

The minimum screen resolution to display the graphical user interface of CiteSpace properly is 1,024 x 768. Versions 4.0.R5 SE or higher should be able to display everything properly at 1,024 x 768. In addition, most functions

are accessible from the menus as well as direct manipulation controls such as sliders or radio buttons on the user interface.

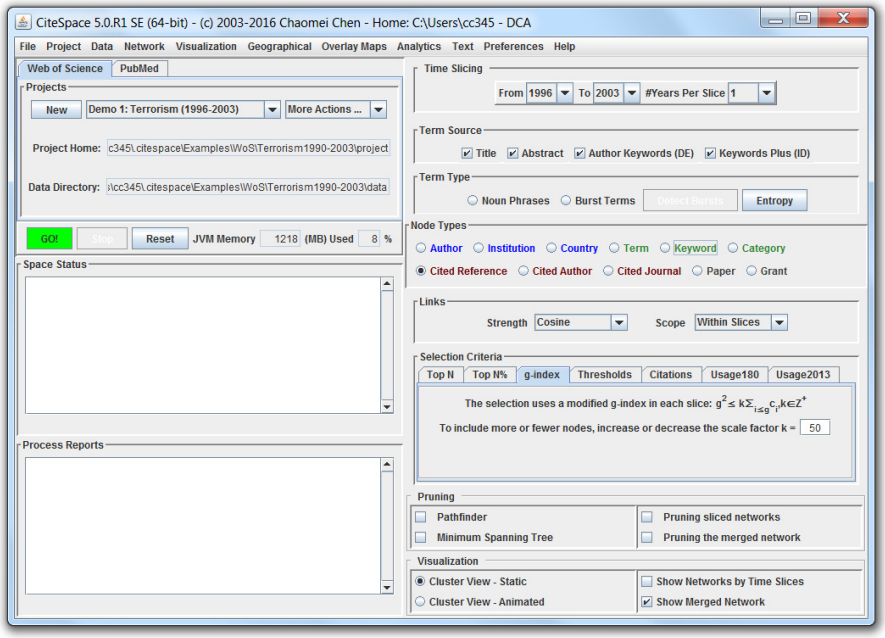

Figure 7. The main interface of CiteSpace.

3.4.1. Geographic Overlays

Geospatial maps depict the distribution of entities of interest on geographic maps. Each entity is marked on a geographic map according to a two-dimensional coordinate of latitude and longitude. It is also possible to position an entity on a three-dimensional coordinate.

3.4.1.1. Geocoding

Geocoding is the process of converting addresses to geographic coordinates. For instance, the address used on my publications is 3141 Chestnut Street, Philadelphia, PA 19104-2875. Using the geocoding service provided by the United States Census Bureau[3], the street address is converted

[3] http://geocoding.geo.census.gov/geocoder/locations/address?form.

to a coordinate in latitude and longitude of (-75.1865, 39.95368). Geographic maps are intuitive and easy to use. We can find nearby researchers in a particular domain or researchers in a vicinity of a specific location.

Figure 8. A collaborative author network on Google Earth.

If an article has multiple authors, then each author is marked on the geographic map. In addition, the connections between them can be depicted on the map by lines joining their locations. If we generate a layer of a geographic representation based on articles published in a particular year, then we can click through different layers to see how geographic patterns change over time.

Multiple layers on a geographic map are a type of overlays in that the geographic map serves as a static base map while changing patterns are revealed by various overlays imposed over the base map. Other types of overlays are possible and base maps do not have to be geographic in nature. For example, celestial and semantic base maps are perfectly acceptable.

3.4.1.2. Generate KML Maps

CiteSpace provides a KML file generator that takes bibliographic records as the input and a thematic overlay of collaborative authors in a KML file as the output. Collaborative relations in each year are represented in lines of different colors.

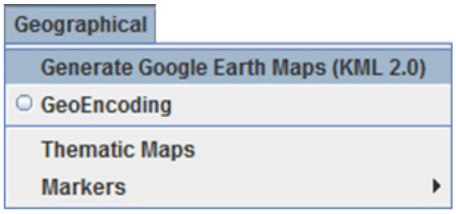

Figure 9. The "Generate Google Earth Maps" function is under the Geographical menu.

To generate the map in KML (a compressed KML format .KMZ) using CiteSpace, first specify the time frame. Articles published outside the range will be ignored. Then, select a folder where bibliographic records in the Web of Science format (plain text) are stored. As we will see later shortly, other functions in CiteSpace deal with the same format. Press the Make Map button to start the process, which parses the data and looks up the latitude and longitude of each address. This process is known as geocoding.

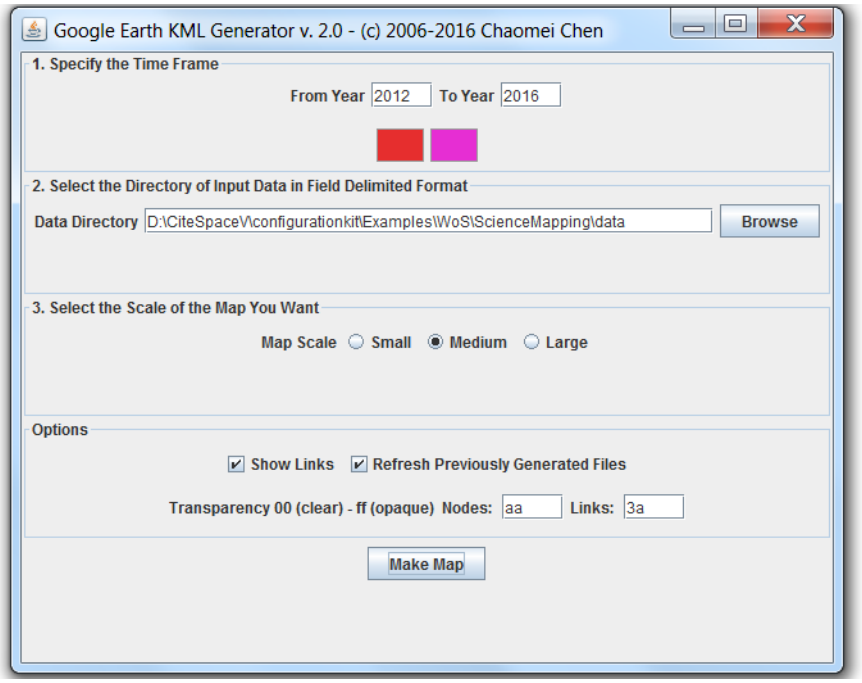

Figure 10. Configure the Google Earth KML Generator. The filenames of data files must begin with 'download.'

The geocoding process may take several minutes, depending on the sample size of your data. The process will create a sub-directory kml in the data folder you specified earlier. Once the process is completed, you will see a dialog box that notifies you of the creation of a kmz file.

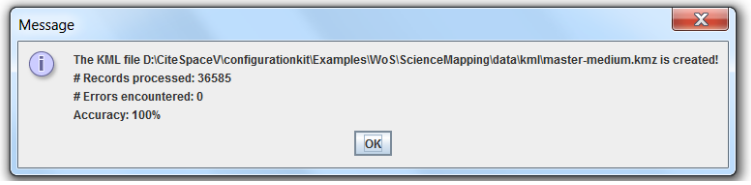

Figure 11. A KMZ file is created.

If the KML generator encounters any errors in the process, relevant information about the errors will be saved to a log file in the kml sub-directory – geocoding_log_tab.txt – so that you may use the information to find a solution. The geographic map is in the file called master-medium.kmz, assuming you selected the medium one as the default scale. In order to view the map in the kmz file, you need to have Google Earth installed on your computer.

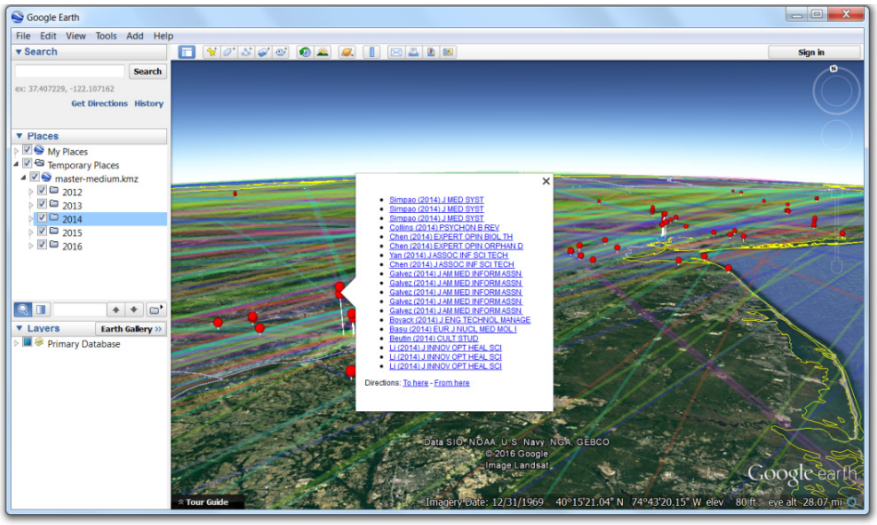

Figure 12. Science mapping related publications in 2014 by authors in our neighborhood.

If Google Earth is installed on your computer, you can double click on the kmz file. In Google Earth, you will see a list of layers on the left, with each layer representing a year of publication. Each layer contains a list of locations on the map where one or more authors published articles in the dataset. You can select or unselect any layer by using the checkbox in front of them.

You will see some vertical bars on the map. The height of a bar indicates the number of publications at the location. Hoovering your mouse cursor over the bar will bring up a list of publications. Each item on the pop-up list is hyperlinked to the webpage of the original publication based on its DOI. If the article is covered by your or your institution's subscription, you should be able to directly access the full text of the article. Coauthor links in more recent years are colored in red, whereas older collaborations are shown in green or blue lines.

Figure 13. The full text of our own 2014 review of regenerative medicine.

In addition to Google Earth, you can use your smart phones to view the generated KML file. Remember that the KMZ file is just a compressed KML file. Uncompress the KMZ file and after several layers of directories you will see the KML file all the way down. Email the KML file to yourself and open it on your smart phone.

3.4.1.3. Geospatial Heatmaps

Given a KML file generated by CiteSpace, creating heatmaps using Google Fusion Tables is fairly straightforward. Google Fusion Table requires the uncompressed KML format, whereas CiteSpace produces the KML file in its compressed version, i.e., the KMZ format. Unzip the KMZ file first and then load the KML file to Google Fusion Table. The following heatmap is an example of the geographic distribution of authors who published articles about Ebola between 1980 and 2014 as of October 19, 2014. This heatmap may give you some ideas of where you may find experts on Ebola.

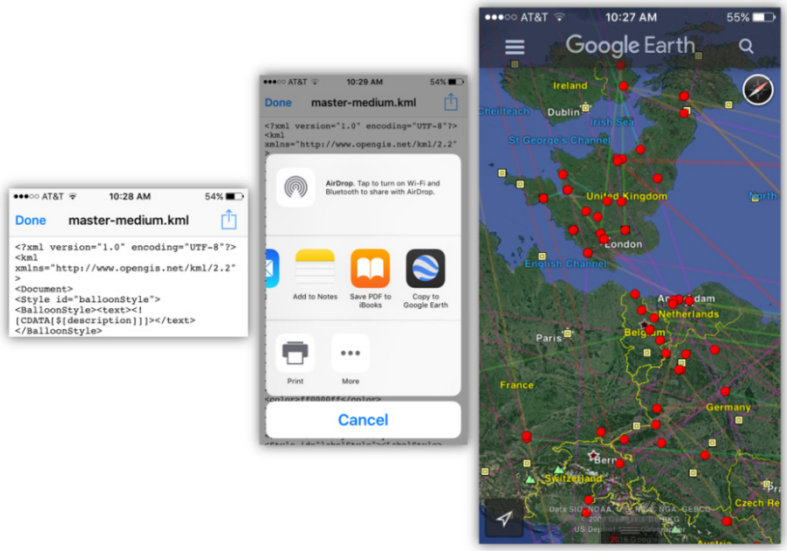

Figure 14. Copy the KML file to Google Earth on iPhone.

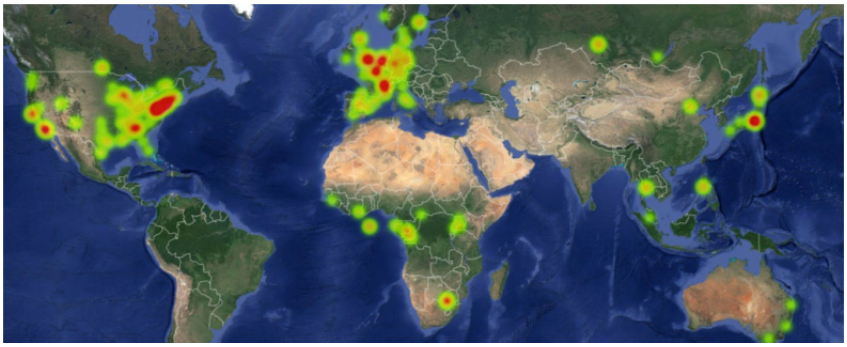

Figure 15. A heatmap of publications on Ebola (1980-2014), as of October 19, 2014.

Here is another example of a heatmap on the impact of retracted papers – the geographic distribution of papers that cited retracted papers.

On January 30, 2014, Haruko Obokata et al., published two papers[4,5] in Nature about STAP – a high-profile event at the time. However, both papers were retracted in July 2014. An eventually retracted paper can be cited before and long after its retraction. A heatmap of authors who cited the retracted STAP papers will provide an overview of who was paying attention to the retracted papers.

The heatmap is created using a KML map generated by CiteSpace. The original source data is from the Web of Science using a search by author for the name of Obokata H*. The search found 4 records. For each of the STAP papers, one record is for the original publication and the other is the retraction notice. According to the citation report, 129 articles on the Web of Science cited the 4 records. Our question is where the authors of these 129 articles are located.

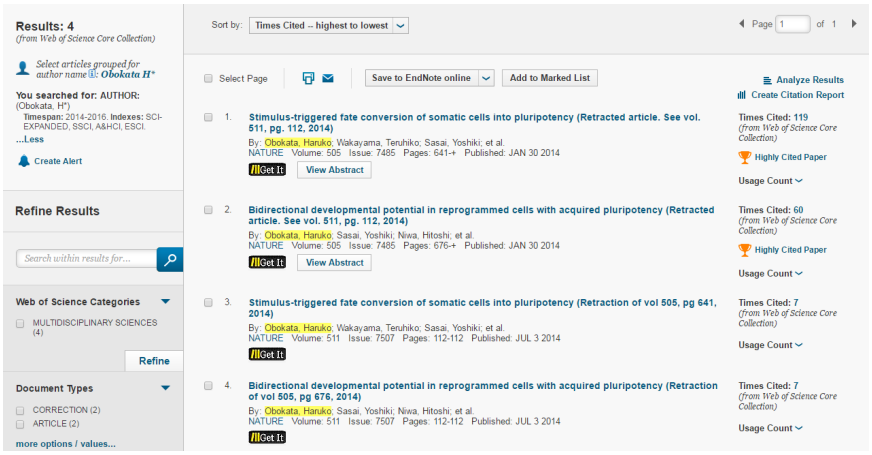

Figure 16. The results of search by author name Obokata, H*.

Heatmaps are intuitive to interpret. Areas with a high density of red indicate cities with high concentrations of authors who cited the STAP papers, including cities on the east coast and the west coast of the United States, the United Kingdom and some part of the Europe, and Japan.

[4] http://www.nature.com/nature/journal/v505/n7485/full/nature12968.html.
[5] http://www.nature.com/nature/journal/v505/n7485/full/nature12969.html.