Analyzing and Visualizing the Dynamics of Scientific Frontiers and Knowledge Diffusion

Chaomei Chen
Drexel University, USA

Natasha Lobo
Drexel University, USA

INTRODUCTION

Estimated numbers of scientific journals in print each year are approximately close to 70,000-80,000 (Rowland, McKnight, & Meadows, 1995). Institute of Scientific Information (ISI) each year adds over 1.3 million new articles and more than 30 million new citations to its science citation databases of 8,500 research journals. The widely available electronic repositories of scientific publications, such as digital libraries, preprint archives, and Web-based citation indexing services, have considerably improved the way articles are being accessed. However, it has become increasingly difficult to see the big picture of science.

Scientific frontiers and longer-term developments of scientific disciplines have been traditionally studied from sociological and philosophical perspectives (Kuhn, 1962; Stewart, 1990). The *scientometrics* community has developed quantitative approaches to the study of science. In this article, we introduce the history and the state of the art associated with the ambitious quest for detecting and tracking the advances of scientific frontiers through quantitative and computational approaches. We first introduce the background of the subject and major developments. We then highlight the key challenges and illustrate the underlying principles with an example.

BACKGROUND

In this section, we briefly review the traditional methods for studying scientific revolutions, and the introduction of quantitative approaches proposed to overcome cumbersome techniques for visualizing these revolutions.

The concept of *scientific revolutions* was defined by Thomas Kuhn in his *Structure of Scientific Revolutions* (Kuhn, 1962). According to Kuhn, science can be characterized by normal science, crisis, and revolutionary phases. A scientific revolution is often characterized by the so-called *paradigm shift*.

Many sociologists and philosophers of science have studied revolutions under this framework, including the continental drift and plate tectonics revolution in geology (Stewart, 1990) and a number of revolutions studied by Kuhn himself. Scientists in many individual disciplines are very interested in understanding revolutions that took place at their doorsteps, for example, the first-hand accounts of periodical mass extinctions (Raup, 1999), and superstring revolutions in physics (Schwarz, 1996).

Traditional methods of studying scientific revolutions, especially sociological and philosophical studies, are time consuming and laborious; they tend to overly rely on investigators' intimate understanding of a scientific discipline to interpret the findings and evidence. The lack of large-scale, comparable, timely, and highly repeatable procedures and tools have severely hindered the widespread adaptation and dissemination such research. Scientists, sociologists, historians, and philosophers need to have readily accessible tools to facilitate increasingly complex and time-consuming tasks of analyzing and monitoring the latest development in their fields.

Quantitative approaches have been proposed for decades, notably in scientometrics, to study science itself by using scientific methods, hence the name science of science (Price, 1965). Many expect that quantitative approaches to the study of science may enable analysts to study the dynamics of science. Information science and computer science have become the major driving forces behind the movement. Commonly used sources of input of such studies include a wide variety of scientific publications in books, periodicals, and conference proceedings. Subject-specific repositories include the ACM Digital Library for computer science, PubMed Central for life sciences, the increasing number of open-access preprint archives such as www.arxiv.org, and the World Wide Web.

THE DYNAMICS OF RESEARCH FRONTS

Derek Price (1965) found that the more recent papers tend to be cited about six times more often than earlier papers. He suggested that scientific literature contains two distinct parts: a *classic* part and a *transient* part, and that the two parts have different *citation half-lives*. Citation half-lives mimic the concept of half-life of atoms, which is the amount of time it takes for half of the atoms in a sample to decay. Simply speaking, classic papers tend to be longer lasting than transient ones in terms of how long their values hold. The extent to which a field is largely classic or largely transient varies widely from field to field; mathematics, for example, is strongly predominated by the classic part, whereas life sciences tend to be highly transient.

The notion of *research fronts* is also introduced by Price as the collection of highly-cited papers that represent the frontiers of science at a particular point of time. He examined citation patterns of scientific papers and identified the significance of the role of a quantitative method for delineating the topography of current scientific literature in understanding the nature of such moving frontiers.

It was Eugene Garfield, the founder of the Institute for Scientific Information (ISI) and the father of *Science Citation Index* (SCI), who introduced the idea of using cited references as an indexing mechanism to improve the understanding of scientific literature. Citation index has provided researchers new ways to grasp the development of science and to cast a glimpse of the big picture of science. A *citation* is an instance of a published article ***a*** made a reference to a published item ***b*** in the literature, be a journal paper, a conference paper, a book, a technical report, or a dissertation. A citation is directional, $\boldsymbol{a} \Rightarrow \boldsymbol{b}$. A *co-citation* is a higher-order instance involving three articles, ***a***, $\boldsymbol{b}_i$, and $\boldsymbol{b}_j$ if we found both $\boldsymbol{a} \Rightarrow \boldsymbol{b}_i$ and $\boldsymbol{a} \Rightarrow \boldsymbol{b}_j$. Articles $\boldsymbol{b}_i$ and $\boldsymbol{b}_j$ are co-cited. A citation network is a directed graph, whereas a co-citation network is an undirected graph.

Researchers have utilized co-citation relationships as a clustering mechanism. As a result, a cluster of articles grouped together by their co-citation connections can be used to represent more evasive phenomena such as specialties, research themes, and research fronts. Much of today's research in co-citation analysis is inspired by Small and Griffith's (1974) work in the 1970s, in which they identified specialties based on co-citation networks. A detailed description of the subject can be found in Chen (2003). A noteworthy service is the ISI Essential Science Indicators (ESI) Special Topics, launched in 2001. It provides citation analyses and commentaries of selected scientific areas that have experienced recent advances or are of special current interest. A new topic is added monthly. Other important methods include *co-word analysis* (Callon, Law, & Rip, 1986). A fine example of combining co-citation and co-word analysis is given by Braam, Moed, and van Raan (1991).

KNOWLEDGE DIFFUSION

Knowledge diffusion is the adaptation of knowledge in a broad range of scientific and engineering research and development. Tracing knowledge diffusion between science and technology is a challenging issue due to the complexity of identifying emerging patterns in a diverse range of possible processes (Chen & Hicks, 2004; Oppenheimer, 2000).

Just as citation indexing to modeling and visualizing scientific frontiers, understanding patent citations is important to the study of knowledge diffusion (Jaffe & Trajtenberg, 2002). There are a number of extensively studied knowledge diffusion, or knowledge spillover, cases, namely liquid crystal display (LCD), and nanotechnology. Knowledge diffusion between basic research and technological innovation is also intrinsically related to the scientific revolution. Carpenter, Cooper, and Narin (1980) found that nearly 90% of journal references made by