Research Policy 43 (2014) 1450–1467

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Identifying emerging topics in science and technology☆

Henry Small[a], Kevin W. Boyack[b,*], Richard Klavans[c]

[a] SciTech Strategies, Inc., 105 Rolling Road, Bala Cynwyd, PA 19004, USA
[b] SciTech Strategies, Inc., 8421 Manuel Cia Pl. NE, Albuquerque, NM 87122, USA
[c] SciTech Strategies, Inc., 2405 White Horse Road, Berwyn, PA 19312, USA

ARTICLE INFO

Article history:
Received 16 April 2013
Received in revised form 21 January 2014
Accepted 14 February 2014
Available online 12 March 2014

Keywords:
Citation-based modeling
Emerging topics
Scientific discovery
Technological innovation
Exogenous events

ABSTRACT

The identification of emerging topics is of current interest to decision makers in both government and industry. Although many case studies present retrospective analyses of emerging topics, few studies actually nominate emerging topics for consideration by decision makers. We present a novel approach to identifying emerging topics in science and technology. Two large scale models of the scientific literature, one based on direct citation, and the other based on co-citation, are combined to nominate emerging topics using a difference function that rewards clusters that are new and growing rapidly. The top 25 emergent topics are identified for each year 2007 through 2010. These topics are classified and characterized in various ways in order to understand the motive forces behind their emergence, whether scientific discovery, technological innovation, or exogenous events. Topics are evaluated by searching for recent major awards associated with the topic or its key researchers. The evidence presented suggests that the methodology nominates a viable list of emerging topics suitable for inspection by decision makers.

1. Introduction

The evolution of topics, including emerging topics in science and technology, has been of interest to governments, companies, and individual scientists for a number of years. Sponsored research in this area has come in waves. For example, in the United States the NSF TRACES program of the 1960s attempted to trace important events in the R&D process. DARPA's Topic Detection and Tracking (TDT) program started in the late 1990s and ran for several years. More recently, IARPA's Foresight and Understanding from Scientific Exposition (FUSE) program (http://www.iarpa.gov/Programs/ia/FUSE/fuse.html) was funded in 2011 to "develop automated methods that aid in the systematic, continuous, and comprehensive assessment of technical emergence using information found in published scientific, technical, and patent literature." The recent America Competes Act explicitly mentions identification of emerging and innovative areas as a specific goal. Today there are conferences and societies dedicated to the study of emerging technologies.

Despite this long-term and recent interest in emerging technologies and its prominence as a topic of interest – a Scopus search for "emerging technology(ies)" returns over 13,000 articles – identification of emerging topics in science and technology remains a challenge. In a recent review of definitions and techniques, Cozzens et al. (2010) report that most studies of emerging technologies are retrospective analyses of pre-determined areas rather than methodological studies designed to identify emerging technologies. For example, Takeda et al. (2009) named nanobiotechnology as an emerging and important domain within nanotechnology, and then used bibliometric techniques to characterize the structure of topics within that domain. While characterization of recent work is important and helps current participants in a technology to understand its history and landscape, these types of studies cannot identify the currently emerging topics that are of interest to funding bodies and practitioners worldwide. Few studies have proposed techniques to identify emerging topics and fewer still have nominated a list of emerging topics from the literature.

This study proposes a new technique for identifying emerging topics from a broad citation database, and uses that technique to nominate over 70 topics from recent years as emergent. These topics are characterized in terms of their key inception events and motive forces, applications and various metrics. Evidence is

☆ A preliminary version of this paper based on 2010 data only was presented at the ISSI 2013 conference (Small et al., 2013).

* Corresponding author. Tel.: +1 505 856 1267.
E-mail addresses: hsmall@mapofscience.com (H. Small), kboyack@mapofscience.com (K.W. Boyack), rklavans@mapofscience.com (R. Klavans).

http://dx.doi.org/10.1016/j.respol.2014.02.005

gathered and presented to show that these topics and their key researchers are extraordinary in many ways, thus suggesting that the methodology produces very useful results. This paper proceeds with a discussion of related work, which is followed by descriptions of the new technique, nominated emergent topics and their characterization, evidence associated with those topics, and a discussion of the results in the context of science policy.

2. Related work

2.1. Defining emergence

The concept of *emergence* is one that is "widely used but seldom defined" (Cozzens et al., 2010), even among studies of emerging technologies. This is perhaps due to the fact that the term *emergence* is used in many different ways (Corning, 2002; De Haan, 2006). As it relates to topics in science and technology, Alexander et al. (2012) provide a history of emergence and its various usages. Goldstein (1999) ascribed the following properties to emergence: radical novelty; coherence, correlation, wholeness; global or macro; dynamical (not pre-given wholes); and ostensive, perceivable. When comparing these properties with those from other definitions, there is nearly universal agreement on two properties associated with emergence – novelty (or newness) and growth.

2.2. Identifying emergent topics

While most retrospective analyses of emerging technologies have been focused simply on characterization of their topics of interest, some few studies have been conducted to develop methods to more easily identify emerging topics. Cozzens et al. (2010) classified these automated methods into two main groups – (1) searching for rapid growth of publications in an existing category or vocabulary (e.g., MeSH) structure, and (2) data mining, which is further characterized as creating structure from a data set using co-occurrence clustering (e.g., co-author, co-word, co-citation) and looking for emergence within that structure. A third group allowed for combinations of these two main approach types.

Of the methods based on identification of rapid growth within categories or vocabularies, the burst detection approach of Kleinberg (2002) is perhaps the most widely used. Kleinberg models time-dependent data using an infinite-state automaton; bursts appear naturally as state transitions. Although originally designed to analyze data streams (e.g., news articles), it has been widely adopted for bibliometrics use (cf., Mane and Börner, 2004) and has been incorporated in larger tool sets such as Citespace II (Chen, 2006), Sci2, and the Network Workbench (Börner et al., 2010). Other studies have used simpler approaches. For example, Ohniwa et al. (2010) identified emerging MeSH terms in seven different five year periods from 1972 to 2006 by calculating an increment rate for each term and time period, defined as the number of times each term occurred during the final two years divided by the number of times the term occurred during the first three years of the time period. Those terms with the highest increment rate are the most emergent terms. The use of five year periods damped out year-to-year fluctuations in the data, leading to a compelling historical view of the ebb and flow of topics. However, the large time window also makes the method less useful from a recency standpoint.

The use of data mining to create structure (through clustering) which is then analyzed for emergent subtopics has also been explored in different ways. Nearly fifty years ago Garfield (Garfield et al., 1964) pioneered the 'historiograph', using direct citation linkages to show the dominant evolutionary pathways within a research topic. Later, clusters of highly co-cited documents that were linked from year to year were used to detect emergence (Small, 1977). Small identified hot fields (i.e., what we might now call emergent topics) as those clusters with a high number of recent papers and a high mean publication year. Although thresholds and normalizations have changed (Small, 1999; Boyack and Klavans, 2014), the basic process of creating annual co-citation clusters and linking those annual clusters into longitudinal strands or threads has changed very little over the past 35 years. Upham and Small (2010) defined research fronts (co-citation clusters) using ISI (now Thomson Web of Science) data from 1999 to 2004 to identify the top 20 emergent topics within that set. Chen and colleagues used a combination of co-citation analysis and burst detection to characterize emerging trends in the fields of mass extinction and terrorism (Chen, 2006), peptic ulcer, gene targeting and string theory (Chen et al., 2009) and regenerative medicine (Chen et al., 2012). They found that the most emergent clusters were typically associated with key articles that experienced not only a burst in citation counts but which also exhibited high betweenness centrality. In other words, these were clusters that were based on key discoveries that effectively bridged two or more existing topics.

Co-citation is not the only clustering approach that has been used to identify emerging topics. Hopcroft et al. (2004) identified several emerging communities using bibliographic coupling with the Citeseer database, comparing clusterings from two different time periods (1990–1998 and 1990–2001). They found that although small changes in the data typically led to significant changes in the clusters, using those few clusters that remained largely unchanged over several clustering runs produced good results. Direct citation, the technique at the core of Garfield's historiography, was later used by Shibata et al. (2008, 2010) to cluster sets of documents on gallium nitride, complex networks, and regenerative medicine. Annual clusterings were done with a fixed starting year – e.g., 1990–2000, 1990–2001, 1990–2002, etc. – and clusters from each model with high overlap were matched and linked to show evolution in field structure. This method is capable of effectively showing births, deaths, splits, and merges in the cluster structure.

Methods that combine growth in vocabularies with rigorous cluster analysis are less common than either of the constituent approaches. Schiebel, Roche and colleagues (Roche et al., 2010; Schiebel et al., 2010) classify keywords from field-based subsets of the PASCAL database as unusual terms, established terms, or cross-section terms based on their relative frequencies. Cluster analysis is used to link clusters of terms between two time periods, and emergent terms are identified as those unusual terms that become established or cross-section in the later time period. Guo et al. (2011) propose a model that simultaneously looks at bursting keywords, growth in number of authors, and changes in the interdisciplinarity of cited references. Results of their study show that emergent areas of science are consistent with a pattern where rapid growth in the number of authors is followed by an increase in the interdisciplinarity of cited references, and then finally by bursts in the keyword structure.

Two relatively recent studies are not easily classified into either of our two main groups of emergence detection methods. Tu and Seng (2012) suggest that the measurement of novelty should be a key part of the identification of novel topics, and define an emergence point at the intersection between a novelty index (1/age) and the cumulative growth curve for a topic. Unfortunately, this method requires sufficient time for the growth curve to be known, and thus cannot be used accurately for recent topics. Bettencourt et al. (2009) use network analysis to show that the collaboration pattern among researchers within an emergent topic experiences a distinct and rapid topological transition from small disconnected graphs to a large connected component.