Available online at www.sciencedirect.com

ScienceDirect

Technological Forecasting & Social Change 74 (2007) 1519–1538

Technological Forecasting and Social Change

Assessment of China's and India's science and technology literature — introduction, background, and approach

Ronald N. Kostoff [a,*], Sujit Bhattacharya [b], Michael Pecht [c]

[a] *Office of Naval Research, 875 N. Randolph St., Arlington, VA 22217, USA*
[b] *National Institute of Science, Technology and Development Studies (NISTADS), Pusa Gate, K.S.Krishnan Marg, New Delhi-110012, India*
[c] *Center for Advanced Life Cycle Engineering (CALCE), University of Maryland, College Park, MD 20742, USA*

Received 9 February 2007; accepted 16 February 2007

Abstract

Science and technology (S&T) allows (1) automation to replace human labor, (2) enhanced human labor capabilities, (3) quicker and cheaper production of goods, and (4) more complex products and processes. In order to maintain competitive advantages, it is critical for any country to understand what other countries are producing in S&T, and what intrinsic S&T capabilities are being developed.

India and China are the two most populous countries in the world. These two dynamic economies are advancing rapidly in S&T, and it is prudent to assess the quantity and quality of their research output as well as to examine trends in their S&T capabilities.

This paper, the first of four in a Special Section on China's and India's S&T, introduces the remaining three papers. Specifically, this paper describes the motivation for the studies, the background for understanding national S&T assessments, an overview of text mining, a brief picture of the Indian and Chinese S&T establishments, and a summary of the analytical techniques used in the assessments.
Published by Elsevier Inc.

Keywords: China; India; Science and technology; Research evaluation; Research assessment; Technology assessment; Text mining; Bibliometrics; Computational linguistics; Document clustering; Factor analysis; Correlation mapping; Core competency

* Corresponding author. Tel.: +1 703 696 4198; fax: +1 703 696 8744.
E-mail address: kostofr@onr.navy.mil (R.N. Kostoff).

0040-1625/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.techfore.2007.02.004

1. Introduction

The present Special Section examines the S&T literature of India and China, the two most populous countries in the world. Both these countries are among the fastest growing economies of the world. Their growth has been attributed to the liberalization of their economies: China in late 1980s and India in the early 1990s, helping these countries to integrate with the world market. This transition has affected these two countries in many ways, such as the changes in composition of their economies from primarily agrarian towards technology governed activities. China has emerged as a leading center for manufacturing industry, whereas India is emerging as a hub for providing services (primarily software development) for the global economy. Both have low cost advantages and large highly-trained manpower pools that add to their competitive strengths. They are striving to assert their presence in the world-market as technologically sophisticated countries [1]. Additionally, India is expanding its capabilities in the manufacturing sector, and China is expanding in the service sector market. Both these countries are demonstrating technological capabilities that have been the forte of technologically developed economies. Thus, India–China studies evoke special interest. The present Special Section undertakes a detailed assessment of the S&T capabilities of these two countries based on research output in journals. The quantity and quality of their research output as well as trends in their S&T capabilities are examined in this context.

Table 1 provides some idea of the intrinsic resources of these emerging economies, using the USA for comparison.

The above table provides some interesting insights into these two countries. China's GDP is more than double that of India, and about 70% that of the USA. India's population is substantially younger than that of China, and China's population in turn is younger than that of the USA. At the same time, India's birth rate is almost double that of China or the USA.

One important indicator is the investment in R&D as percentage of GDP. Until 1998, India's R&D intensity was higher then that of China, in the range of 0.6 to 0.8%. However, China has now reached 1.4% level whereas India is still below the 1% level. Another important indicator is Foreign Direct Investment (FDI) in R&D. Both countries are emerging as favorable locations of foreign R&D centers. A United Nations Conference on Trade and Development (UNCTAD) survey [2] during 2004–2005 of the world's largest R&D spenders shows the growing importance of these two countries as R&D locations. About 35% of the major Transnational Corporations (TNCs) already have R&D centers in China, whereas about 25% have centers in India. Moreover, China is the destination mentioned by the largest number of respondents for future R&D expansion followed by the United States. Third place in the global destination choice of

Table 1
Comparison of India–China–USA resources

India–China–USA resource comparison (2006)

Resource	India	China	USA
Land mass (M SQ KM)	3.29	9.6	9.63
Population (B)	1.1	1.31	.298
Med age (years)	24.9	32.7	36.5
Birth rate/1000	22	13.3	14.1
GDP (T): (purchasing power parity)	3.61	8.86	12.36
GDP (T): (official exchange rate)	0.72	2.23	12.49

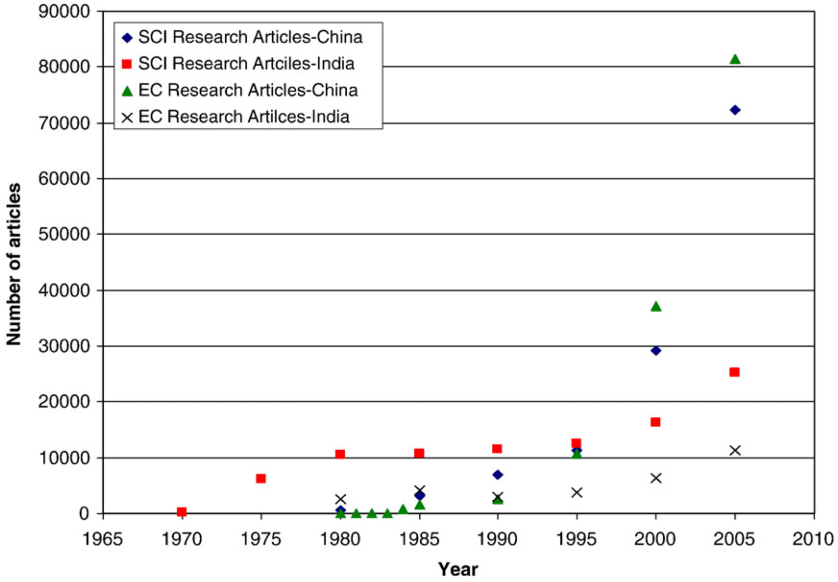

Fig. 1. Number of research articles by Chinese or Indian authors.

foreign R&D locations is India. Other important reports such as The Economist's World Investment prospect 2004 [3] highlight the attractiveness of these two countries as foreign R&D locations.

In recent years, both India and China have expanded their efforts in S&T. Fig. 1 compares the number of research articles by Chinese and Indian authors over the last four decades.[1] Prior to 1980, Indian publications greatly outpaced those by the Chinese.

In 1980 – about the time Chinese articles began to appear in the databases – Chinese publication rates began an exponential rise of about 20% per year. Over the period 1980–1995, Indian publications remained more or less constant. Indian and Chinese publications were about equal in 1995. From 1995 to 2005, the growth in Chinese publications has greatly exceeded that of India's.

The series of papers in this Special Section attempts to answer some critical questions using suites of tools and processes known collectively as text mining. The key critical questions related to the S&T research literature focus on quality, technical scope, areas of strategic emphasis, core competencies and critical mass, key performers, significant external collaborators, the sources of cutting-edge research, research citations, and India's and China's awareness of external S&T.

There are four papers in this Special Section. In the present paper, we describe the motivation for the studies, the background for understanding national S&T assessments, an overview of text mining, a brief picture of the Indian and Chinese S&T establishments, and a summary of the analytical techniques used in the assessments. The second paper presents an assessment of China's S&T literature, and the third presents an assessment of India's S&T literature. The fourth paper compares India's and China's research outputs from multiple perspectives.

[1] The data were taken from the Science Citation Index/ Social Science Citation Index (SCI/ SSCI) and the Engineering Compendex (EC).

2. Background

Three essential factors are explored in these papers: S&T assessments by country, text mining, and the current state of research in India and China. The first segment of this section presents a brief historical survey of country S&T assessments to provide proper perspective. The second segment presents an overview of text mining, focusing on its major bibliometrics and computational linguistics components to provide an understanding of why these particular instruments were selected. The third section provides a very brief overview of China's and India's S&T establishments to place the thematic research outputs in context.

2.1. Country technology assessments

Country technology assessments aim to identify the range and breadth of technologies sponsored and conducted by the country of interest. These assessments emphasize the technical infrastructure associated with the diverse technologies at different levels of aggregation (key performers, centers of excellence, etc.), and identification of the national research core competencies. What do we mean by national research core competencies, and why are they important?

A national research core competency is a synergy of individual expertise that is aggregated and coordinated over multiple technical disciplines and expressed as a national research strategic investment. More specifically, a national research core competency is a technical area that (1) engages a critical mass of researchers; (2) consists of coordinated and synchronized sub-disciplines; (3) produces high-quality output; (4) offers unique national capabilities; and (5) contains a visible fraction of research investment ([4,5]). National S&T core competencies represent a country's strategic capabilities in S&T.

The next two papers in this Special Section present a text-mining approach that addresses a sub-set of China's and India's national research core competencies in terms of the identification of their main research thrusts, volume of research output in those areas, and relative quality of selected major research thrusts. Further subjective analysis (beyond the scope of the present paper) is required to characterize the remaining necessary features of a national core competency. Knowledge of country core competencies is important to be able to perform the following:

(a) prioritizing technical areas for joint commercial or military ventures,
(b) assessing a country's military potential, and
(c) understanding emerging areas to avoid commercial or military surprises.

Obtaining such global technical awareness, especially from the literature, is difficult for several reasons:

(a) Much of the science and technology performed is not documented.
(b) The science and technology that is documented is not widely available.
(c) The available documented science and technology is expensive and difficult to acquire.
d) Few credible techniques exist for extracting useful information from large amounts of science and technology documentation ([6]).

Most credible country technology assessments are based on a combination of personal visits to the country of interest, supplemented by copious reading of technology reports from that country. Such processes tend to be laborious, slow, expensive, and accompanied by large gaps in the available knowledge.

Over the past half century, driven mainly by the Cold War, a large number of country technology assessments were performed ([7–18]). The last decade has seen an expansion in focus to technologies of major economic competitors. Over the last two decades, some of the most credible of these assessments have come from two organizations: the World Technology Evaluation Center (WTEC) at Loyola University and the Foreign Applied Sciences Assessment Center (FASAC-SAIC). In conducting their studies, both of these organizations gather topical literature from the country of interest, assemble teams of experts in the topical area that review the literature and conduct site visits, and ask the teams to brief their findings and write a final report. The studies performed by these groups remain seminal approaches to country technology assessments. Some recent studies of China focusing on bibliometrics and using network approaches have been performed, and they offer the insights only a comprehensive computer-based analysis can provide ([19,20]).

2.2. Text-mining technology assessments

Text-mining approaches have been developed to extract useful information from the global science and technology literature for the past decade ([21–28]). These studies have typically focused on a technical discipline, and have examined global S&T efforts in this discipline. Such approaches, with slight modification, could be adapted to identify core S&T competencies in selected countries or regions, including estimation of the relative levels of effort in each of the core technology areas. Moreover, coupling the text-mining approach with the WTEC and FASAC approaches would amplify the strengths of each approach and reduce its limitations.

Once the key technologies, researchers, and Centers of Excellence have been identified through text mining, site visitation strategies can be developed. The second phase of the effort would be actual site visitations. A key step in this hybrid process would be the demonstration of the ability of text mining to identify targets of interest with reasonable precision in a timely manner at an acceptable cost. These three driving parameters (performance, time, cost) could be traded off against each other to provide a balance acceptable and tailored to a variety of potential customers.

A typical text-mining study of the published literature develops a query for comprehensive information retrieval, processes the retrieved database using computational linguistics and bibliometrics, and integrates the processed information. Science and technology computational linguistics [6] identifies pervasive technical themes in large databases from technical phrases that occur frequently. It also identifies relationships among these themes by grouping (clustering) these phrases (or their parent documents) on the basis of similarity. Computational linguistics can be used for:

- enhancing information retrieval and increasing awareness of the global technical literature [29–31];
- potential discovery and innovation based on merging common linkages among very disparate literatures [32–34];
- uncovering unexpected asymmetries from the technical literature (For example, Kostoff predicted asymmetries in recorded bilateral organ (lungs, kidneys, testes, ovaries) cancer incidence rates from the asymmetric occurrence of lateral word frequencies (left, right) in Medline case study articles.) [35,36];
- estimating global levels of effort in S&T subdisciplines [27];
- helping authors potentially increase their citation statistics by improving access to their published papers, thereby potentially helping journals to increase their impact factors [27]; and
- tracking myriad research impacts across time and applications areas [26].

Evaluative bibliometrics [37,38] uses counts of publications, patents, citations, and other information to develop science and technology performance indicators. Its validity is based on the premises that (1) numbers of patents and papers provide valid indicators of R&D activity in their subject areas; (2) the frequencies with which patents or papers are cited in subsequent patents or papers provide valid indicators of the impact or importance of the cited works; and (3) the citations from papers to papers, from patents to patents, and from patents to papers provide indicators of intellectual linkages among the organizations producing the materials, and knowledge linkages among their subject areas [39]. Evaluative bibliometrics can be used to provide the following functions:

- identify the infrastructure (authors, journals, institutions) of a technical domain;
- identify experts for innovation-enhancing technical workshops and review panels;
- develop site visitation strategies for assessment of prolific organizations globally; and
- identify impact (literature citations) of individuals, research units, organizations, and countries.

Understanding assessment methodologies is one part of the equation. Understanding the technical organization of the countries to be assessed is equally important.

2.3. China's science and technology enterprise

China regards basic research as the foundation for the development of future technologies, as well as a driving force for sustainable long-term development of its economy [40–42]. As a developing country, China has adopted an S&T development policy requiring that available resources be concentrated on the development of selected high technologies critical to the nation's economic development. In fact, similar strategies have been applied to many other government-funded development programs, such as China's military modernization programs ([43]). Strengthening basic research has been a goal during the Ninth, and now the Tenth, Five-year Plan (FYP) periods. Both FYPs called for efforts to make breakthroughs in selected areas ([44]).

Since 1997–1998, China's gross expenditure on Research and Development (GERD) growth has been slightly higher than its gross domestic product (GDP) growth, reflecting the government's accelerated effort in S&T development. China has been encouraging product-development R&D activities to ensure S&T contributes to its economic development. For example, in 2002, 75% of the nation's R&D spending went to product development and another 19% to applied research ([45]). In 2002, the Chinese Academy of Science (CAS) increased its spending on basic research to 40% of its total outlay, aiming at Nobel-level fundamental research. It has also taken measures to increase its scientists' creativity ([46]).

Despite this, many Chinese scientists argue that basic research is seriously underfunded. In 2001, China's basic research funding in the country was 5.3% of total R&D expenditures, compared with a ratio of 16 to 20% in the United States, Western Europe, and Japan ([47]). In 2003, China had about 0.86 million people involved in R&D activities, compared with 1.26 million in the U.S. and about 0.67 million in Japan ([48]). China's R&D spending remains at a low level in terms of the GERD–GDP ratio as compared with several scientifically important developed countries, and this situation is unlikely to change significantly in the near future. Among seven large OECD economies with population over 40 million, six (the U.S., Japan, U.K., Germany, France, and South Korea) exhibit current R&D ratio (R&D/GDP) in the range of 2 to 3%. Italy is the only exception and has yet to attain this level. Singapore has recently made the transition from that of a low-intensity R&D country (<1%) to a high intensity R&D

country (>2%). On average, the six OECD economies required 10 years to make the transition from 1% to level off in the 2–3% range [49]. China's R&D intensity exhibits a continuous upward trend, crossing the 1% threshold in 2000 and rising to 1.4% in 2004. The intensity has more then doubled from 1995 when it was only 0.6%. China's GDP has also significantly risen during this period and thus it can be concluded that there has been significant investment in R&D to attain the present intensity. The upward trend plausibly indicates that China is in the transition phase and would eventually attain the R&D intensity in the 2–3% range.

In 2004, state-owned enterprises accounted for 66.83% of the total R&D performed in the country, R&D institutes for 21.95%, and universities for 10.22% ([44]). China (like most developed scientific countries, including the United States and Japan) also encourages non-governmental sectors to support R&D with their own funds. In 2003, governments (central and provincial) contributed 29.9% of total R&D support in China, enterprises 60.1%, foreign sources 2%, and the remaining 8% was accounted for by unspecified "other" sources. However, among the enterprises' expenditures, it was estimated that approximately half the R&D funding came from state-owned enterprises (SOEs), and thus indirectly from the central government. If so, then 62% of China's R&D expenditures in 2004 came either directly or indirectly from government and only 29% purely from private enterprises. In the United States, private industry accounts for over 65% of all R&D support, with government accounting for somewhat less than 30%. In Japan, private industry accounts for a slightly higher percentage of total R&D support than in the United States, and government for slightly less ([50]).

The State Council of the central government is the highest administrative body of China. There are six major ministry-level administrative organizations directly under the State Council that handle the nation's S&T development activities. These organizations include the Ministry of Science and Technology (MOST), the Ministry of Education (MOE), the Commission of Science, Technology and Industry for National Defense (COSTIND), the Chinese Academy of Sciences (CAS), the Chinese Academy of Engineering (CAE), and the National Natural Science Foundation of China (NSFC) ([46]). Among those organizations, MOST, COSTIND, and MOE have policy-making authority, in addition to varying degrees of funding authority; CAS (which receives substantial funds from the government as a budget line item to support its research activities) and CAE have advisory power; and NSFC provides research funds. The Leading Group on Science and Technology, chaired by the Prime Minister, is located organizationally between the State Council and these administrative organizations, but most observers agree that it is relatively ineffective in setting R&D priorities.

2.4. India's science and technology enterprise

India has a complex and multi-layered system of science and science administration consisting of governmental agencies, autonomous institutions, the university system and industrial R&D, both in the public and private sectors. Broadly, the S&T system in India can be classified under the following structures: central (federal) government S&T departments/agencies, state (provincial) government S&T departments, central socio-economic ministries, in-house R&D in private industry, S&T in non-governmental organizations (NGOs), and independent research institutes. Central government S&T departments/agencies are the main instruments for providing resources and defining priorities, and are responsible for reaching S&T targets in different sectors. There are twelve scientific departments/agencies mainly involved in R&D activity ([51]). The main functions of these agencies are to support and coordinate research in their respective areas. This is carried out through a chain of laboratories/research

Table 2
Scientific agencies/no. of institutes

Scientific agency	No. of institutes
Indian Council of Agriculture Research (ICAR)	84
Defence Research and Development Organisation (DRDO)	53
Council for Scientific and Industrial Research (CSIR)	38
Indian Council for Medical Research (ICMR)	27
Department of Science and Technology (DST)	17
Department of Atomic Energy (DAE)	14
Department of Electronics (DOE)[a]	14
Department of Space (DOS)	8
Department of Biotechnology (DBT)	5
Department of Ocean Development (DOD)	–
Ministry on Non-Conventional Energy Sources (MNES)	–
Ministry of Information Technology (MIT)	–
Ministry of Environment (MOEn)	–

[a] The Department of Electronics (DOE) is not denoted by DST as a major agency involved in R&D.

institutions that answer to these agencies, as well as through research grants/sponsored projects for the higher education sector, national laboratories, and establishments. The scientific agencies/departments and laboratories under them are shown in Table 2.

The five major scientific agencies – DRDO (Defence Research and Development Organisation), DOS (Department of Space), ICAR (Indian Council of Agricultural Research), DAE (Department of Atomic Energy), and DSIR (Department of Scientific and Industrial Research) (major funding directed to CSIR—Council of Scientific and Industrial Research) – alone accounted for 86.7% of the total R&D expenditures of the central government on scientific agencies or departments in 2002–2003. Maximum priority was given to DRDO, as it received 30.3% of the overall R&D budget. This pattern has not changed significantly in the current period per available projections.

In addition, there are research institutes contributing to research and development under other government ministries for steel, power, railways, and so on (545 institutes, including 285 laboratories under the 12 scientific departments); as state public-sector industrial in-house R&D units; as in-house R&D units of private-sector and non-profit research institutions (1351 unit); and in universities (225 universities)., In all, 2899 institutions are estimated to be carrying out R&D activities in India.[2] Expenditures on R&D by field of science are shown in Table 3.

It is striking to note the low levels of investment in medical sciences across the varying entities. Different priorities can be observed in terms of funding in engineering and technology and agricultural sciences by the central and state governments.

Actual expenditures by ministries and departments include, along with planned expenditures, non-planned expenditures and extramural funding. Thus, it is useful to observe the actual expenditures of each department to accurately assess research funding. The expenditures of thirteen scientific departments/organisations for the period 2002–2003 and 2003–2004 are shown in Table 4.

[2] These figures are from S&T Data Book 2000 (Ministry of Science and Technology, Government of India).

Table 3
Expenditure on research and development by field of science (2002–2003)

Filed of science	Central government	State governments	Public sector	Private sector	Total
Natural sciences	296609.31 (24.68%)	5654.48 (3.69%)	13816.28 (17.07%)	143424.77 (39.31%)	459504.85 (25.53%)
Engineering and technology	734217.48 (61.11%)	1861.92 (1.22%)	66995.85 (82.81%)	179281.34 (49.14%)	982356.58 (54.57%)
Medical sciences	25030.65 (2.08%)	3307.39 (2.16%)	81.87 (0.10%)	29021.22 (7.95%)	57441.13 (3.19%)
Agricultural sciences	145600.56 (12.12%)	142015.21 (92.91%)	0.00 (0.00%)	13097.67 (3.59%)	300713.44 (16.70%)
Total	1201458.00	152839.00	80894.00	364825.00	1800016.00

Source: Research and Development Statistics, 2004–2005.
(Rs. Lakhs)*.
Note: *1 lakh = 0.1 Million.
R&D Percentages are relative to the total expenditure in each sector. All percentages are rounded off.

The high level of funding for the Department of Atomic Energy in comparison with other scientific agencies is evident.

3. Approach

Text mining has two major components for country assessments: infrastructure determination (bibliometrics) and technical structure determination (computational linguistics). The use of these

Table 4
S&T expenditures by various ministries/department/organization

S. no	Ministry/Department/Organization	2002–2003	2003–2004
1	Atomic Energy	6018.73	6148.41
2	Space	2162.22	2268.20
3	Indian Council of Agricultural Research	1333.96	1464.17
4	Scientific and Industrial Research (including grants given to Council of Scientific and Industrial Research)	936.71	1090.09
5	Environment and Forests (including Zoological Survey of India and Botanical Survey of India)	1057.52	1036.19
6	Science and Technology including Survey of India and India Metrological Department	920.84	985.84
7	Information Technology	497.34	530.62
8	Non-Conventional Energy sources	428.33	381.33
9	Geological Survey of India (Ministry of Mines) [a]	248.31	271.60
10	Biotechnology	220.70	262.55
11	Indian Council of Medical Research	180.00	201.86
12	Ocean Development	167.05	169.50
13	Centre for Development of Telemetics (Department of Telecommunications) [a]	108.80	47.66
		14307.51	14858.62

(Rupees in crore); 1 Crore = 10 Million.
Source: CAG (Comptroller and Auditor General) Report: Report No. 5 of 2005 (Scientific Departments).
[a] These are not identified as major scientific agencies by DST-R&D Statistics (2004–2005). Expenditure of the DRDO (Defence Research and Development Organisation) is not covered by the CAG report.

techniques depends on the specific databases involved. Descriptions of the databases employed for the India–China studies and the bibliometrics and computational linguistics techniques used follow.

3.1. Databases and information retrieval approach

The objectives for these studies were to assess the research performances of China and India through analyses of their research output literatures. In the techniques we use for assessment (bibliometrics), both publications and citations play an important role. In our estimation, the premier research output database for both publication and citation bibliometrics is the Science Citation Index/Social Science Citation Index (SCI/SSCI). To access applied literature, the Engineering Compendex (EC) is the database of choice, although it does not have citation capability.

The SCI/SSCI database ([52]) and the EC ([53]) were used as primary sources. The retrieved SCI/SSCI database used for analysis consisted of selected journal records (including the fields of authors, titles, journals, author addresses, author keywords, abstract narratives, and references cited for each paper) obtained by searching the Web version of the SCI for articles that contained at least one author with an address in the country of interest. At the time the final data was extracted for the computational linguistics analysis, the version of the SCI used accessed about fifty-six hundred journals (mainly in physical, engineering, and life sciences basic research), and the version of the EC used accessed about five thousand journals (mainly in applied research, technology development, and engineering).

For the China study, sample records were extracted from the SCI for two different years, 2002 and 2005, and from the EC for years 2000–2003. There were 7780 records with abstracts retrieved from the SCI for 2002, 34,834 records with abstracts retrieved from the SCI for 2004 to early 2005, and 9949 records with abstracts retrieved from the EC for 2000–2003. The abstracts were used for the computational linguistics phase (phrase analyses, document clustering). For the aggregate China bibliometrics analysis, sample 2004–early 2005 records were extracted for publications and sample 2002 records for citations. For the aggregate China bibliometric trend analyses, the total 2005 records were used directly from the SCI, and the SCI Analyze function was used to extract the data. Finally, for the USA–China research investment allocation comparison, approximately ten thousand records were retrieved for each country from the SCI in mid-2006.

For the full India study, records were retrieved from the SCI for 1991, 2002, and 2005. For the present India study described in this Special Section, sample records were extracted from 2005 only. Taxonomy results were obtained for those samples. Fifteen thousand records were retrieved for 2005 (slightly more than half of India's total SCI/SSCI research article output), of which 14308 contained abstracts. The abstracts were used for the computational linguistics. Bibliometrics were performed on the sample 2005 retrieval for mapping and other analyses, and on the total annual records using the SCI analyze function to obtain gross bibliometric trends with time.

For both studies, temporal trend analyses were performed using the analysis capabilities of the SCI and EC display screens. Data from the early 1980s to the present were accessed in both databases in periodic intervals.

3.2. Publication bibliometrics

3.2.1. Trend bibliometrics

An important feature of these studies was examination of the trends of bibliometric quantities over time. Such studies show the temporal evolution of the technical infrastructure, and can identify important trends that might not be obvious in point studies.

Both the SCI and the EC have output display capabilities for computing trends in selected bibliometric quantities. A typical trend analysis consisted of entering a country's name in the address field, displaying the retrieved results on the screen, then applying the analysis tool to the retrieval. No downloading was required for these analyses. Typically, the time origin for analysis was selected as 1980, and then bibliometrics were examined every five or 10 years until 2005.

For most bibliometric computations, a China paper or India paper was defined as a research article having at least one author with a Chinese or Indian address. In some select cases, a China or India paper would be restricted to research articles with only Chinese or Indian author addresses in the address field.

3.2.2. Impact factors

The journal impact factor is a measure of the citation performance of research articles published in the journal. A journal impact factor for 2005, for example, would be the ratio of all 2005 citations of articles published in 2004 and 2003 to the number of articles published in those years. The measure is discipline-dependent, since some disciplines have many more active researchers than others, and these researchers can generate many citations. The impact factor has come to be interpreted as a measure of a journal's quality. Although this is not a linear relationship, the highly regarded journals tend to have relatively high impact factors for their disciplines. The impact factor was used in the present study to provide balance to publication statistics that focused on quantity only.

As will be shown in the specific India and China studies, many of the domestic Chinese or Indian journals had low impact factors. There are many causes that can contribute to a low journal impact factor. These include low quality of publications, limited journal circulation, overly applied papers, and/or size of technical field covered (i.e., number of researchers working in the field and available to cite papers). This study did not distinguish among these factors for the journals listed.

3.2.3. SCI/SSCI access date

Publication of a journal is no guarantee that it will be accessed/indexed by the SCI/SSCI. When a journal becomes sufficiently recognized, it is then accessed by the SCI/SSCI. For some studies, the date at which the journal becomes accessed by the SCI/SSCI may be significant. Why is this initial SCI/SSCI access date important? When evaluating an organization's or country's growth rate in publications (a measure of the research health of the organization or country), two causes of growth are usually included. One could be increased research sponsorship (more funds leading to more papers); the other could be increased research productivity. However, during the period that growth is being examined, if journals are initially accessed by the SCI/SSCI, any papers published in these newly accessed journals will be counted toward publication growth. This apparent research output growth is a bookkeeping artifice. If, over the period of interest, a country publishes extensively in journals newly accessed to the SCI/SSCI, then serious caveats are required before interpreting these "virtual" growth results. This fact is not solely of academic interest. As will be shown in later papers of this Special Section, it is a real concern with regard to China, which publishes extensively in recently accessed journals with low impact factors. This issue is not easily resolvable; there are no easy corrections to be made. The journal may be a good one that should have been accessed by the SCI/SSCI decades earlier; perhaps the country's publications growth should have been reflected at a higher level decades earlier.

3.2.4. Factor analysis

Factor analysis of a database aims to reduce the number of variables in a system and to detect structure in the relationships among variables. Correlations among variables are computed, and highly correlated groups (factors) are identified. The relationships of these variables to the resultant factors are displayed clearly in the factor matrix, in which the rows are variables and the columns are factors. In the factor matrix, the matrix elements, $\mathbf{M}_{ij,}$ are the factor loadings, or the contribution of variable i (in row i) to the theme of factor j (in column j). The theme of each factor is determined by those variables that have the largest values of factor loading. Each factor has a positive value tail and a negative value tail. For each factor, one of the tails typically dominates in terms of absolute value magnitude. This dominant tail is used to determine the central theme of each factor. Factor analysis was used to quantify word/phrase, institution, and country collaborations. The number of factors used in the computations must be specified beforehand, and a number of different approaches exist for selecting this number ([54–56]).

The word/phrase factor matrix is described in the factor analysis sub-section of the computational linguistics section. The infrastructure factor matrices (where the variables can be people, institutions, countries, etc.) identify the main infrastructure groups in a database, essentially based on their record co-occurrences. Identifying the specific infrastructure groups is very straightforward; interpreting why certain infrastructure elements are grouped together is more challenging.

3.2.5. Correlation mapping

An auto-correlation function describes the correlation between a random function and a copy of itself shifted by some ‘lag’ distance. This can be reproduced in a map showing terms that commonly occur together. For example, an auto-correlation map of institutions shows teams of institutions that publish together.

A cross-correlation map shows relationships among items in a list based on the values in another list. For instance, a cross-correlation map of institutions and phrases can show groups of organizations that write about the same things. A cross-correlation map of countries and phrases can show groups of nations that write about the same things.

In the China and India papers, infrastructure auto-correlation maps are compared to cross-correlation maps where possible. This may prove to be a powerful approach for identifying infrastructure elements that are related by common interests but that are not collaborating. The question then is: Why are these collaborations not occurring? Is the cause insufficient awareness, insufficient coordination, or insufficient technical commonality at the lowest levels of technical detail?

Because of the rather modest space available in a paper for a map, care must be taken to extract maximal information from the variables mapped. For example, in cross-correlation maps, the selection of the number of variables to be displayed on the map and the number of variables to be selected for the correlation must be carefully considered.

Typically, the variables selected for display tend to be those of highest frequency, and only those of a sufficient number that they can be displayed clearly. This can be determined iteratively, where extra variables are plotted and those that have little correlation/connectivity on the map are discarded and the correlation re-run. The variables selected for the correlation (those responsible for linking the variables displayed) are more challenging.

Two approaches have been used for selecting specific variables in the India and China papers. In the first, the highest frequency variables are selected for the cross-correlation, and the maps are plotted. Sometimes (not always), a very complex, dense, and highly connected network will be displayed. Because of the line density, very little information is available from the maps. As the first author's ongoing nanotechnology study is showing, cross-correlation plots of performing institutions with cited journals sometimes have this densely connected characteristic (basically, in the institution-cited journal cross-correlation map, such a dense network reflects that most of the institutions are citing the same high-profile journals).

In the second approach, the plotted variables are kept the same, but the correlation variables are selected from lower on the frequency list. One characteristic of bibliometric variables or text phrases is that the variables/phrases become more focused as the frequencies of occurrence decrease. Text phrases become more descriptive technically; cited journals become more discipline-specific. For the cited journals case, the maps are transformed from a highly connected uninterpretable display to a moderately connected display with sharp results (differentiating those institutions that cite journals in very focused technical areas). In the present country studies, different frequency bands were selected for phrases only. Selection of different frequency bands appears to be very promising for text-mining analysis, though this capability does not appear in the literature. The auto-correlation mapping usually precedes the factor analysis, and the number of factors to be selected for the factor analysis is estimated based on the number of groups discernible from the auto-correlation map.

3.3. Citation bibliometrics

3.3.1. Value of collaborations

The ease of international travel and the ready availability of global high-speed communications have resulted in the internationalization of research. Additionally, large centralized facilities created by consortia of countries/international research foundations, as well as projects that require international participation, are other important factors contributing to internationalization of research. This translates into higher levels of research collaboration. To assess the benefits from collaboration for India and China, it is necessary to determine the extent of the collaboration and some measure of its impact.

Country co-occurrence matrices were used to determine the degree to which nations were co-authoring papers with India or China. To quantify the impact of collaboration, citations from articles with and without external collaborators were evaluated. Again, multiple citation perspectives were examined. The first case (including external collaborators) consisted of all research articles in the SCI/SSCI published from 1995 to 1999 having at least one author with a Chinese or Indian address. The second case (no external collaborators) consisted of all research articles in the SCI/SSCI in this period with Chinese or Indian authors only.

The total articles retrieved in the two cases (with and without collaboration) were compared to determine the extent of collaboration. Then, the median citations of the top ten articles, the top 5% articles, and total articles were tabulated and compared. This allowed comparison of India and China from different citation perspectives, as well as each country alone with and without collaboration.

3.3.2. Most-least cited

A powerful technique used in past text-mining studies by the first author is comparison of the most frequently cited papers in a country study or technology assessment with the least cited. This allows infrastructure and technical theme patterns of each literature to be strongly contrasted, providing for easier interpretation.

For both studies in this series, the papers with the most citations (greater than some threshold) were compared with those with the least citations (zero) for separate time periods, in order to track the changes in characteristics of these papers. The most-cited papers were identified visually using the citation sort capability of the SCI operating on the retrieved country records. The papers with zero citations were obtained by random sampling of all articles with zero citations published in the same time frame as the most-cited articles.

3.3.3. Quality comparison using citations

Many of the bibliometric comparisons between India and China have been based on output quantity. But output quality needs to be addressed as well. Comparison of research quality by approaches other than peer review is very complex ([57,58]). We decided to develop a citation-based approach for comparing the quality of China's published research output with that of India. Additionally, for normalization purposes, it was desirable to compare the citation performance of the China–India "winner" with that of a more developed nation. Australia was chosen as a more technologically developed nation located in a similar geographical region (Western Pacific), with a much smaller population and a similar research output for 1998, and was used as a second basis for comparison.

For the comparison, 1998 was chosen as the vintage year. It was long enough ago that a substantial number of citations could have had time to accumulate, but sufficiently recent to indicate current research quality. Additionally, the total number of SCI/SSCI papers for each country for 1998 was similar (India, 16,228 research articles; Australia, 20,185 research articles; China, 18,830 research articles). Equal numbers of records for India, China, and Australia (3500) were downloaded from the SCI/SSCI. Phrases and their frequencies were extracted from each country's materials. China's and India's phrases were combined to compare these two countries, and China's and Australia's phrases were combined to compare these two countries. Identical phrases were grouped, and their ratios of frequency were computed.

It was useful to select phrases representing important technical disciplines with similar levels of emphasis, and since the total published records for each country for 1998 in SCI/SSCI was within about 10%, a factor of about two for the difference in phrase frequency for a technical discipline was viewed as the outer bound of similar emphasis. Thus, those phrases with both high frequencies of occurrence and frequency ratios within a factor of two were extracted, and examined.

For the China–India comparison, different phrases were chosen to represent the four major research categories: physical sciences, environmental/agricultural sciences, life sciences, and materials sciences. Ordinarily, engineering sciences is used rather than materials sciences, but there were insufficient phrases with adequate frequencies to represent engineering sciences, so materials sciences was used instead.

For the China–Australia comparison, different phrases were chosen to represent the same four major research categories.. Each phrase could be perceived as representing a specific technical discipline within one of the four broader categories defined above. Each phrase was used as a separate query, and inserted in the SCI/SSCI search engine for 1998. The total SCI/SSCI citations for the retrieved records for each country for each phrase from 1998 to mid-2005 were tabulated and analyzed.

The philosophy behind the specific metrics used for the comparison is as follows. There are a number of different metrics that could be selected for citation comparisons between the two countries. Average citations, median citations, or citation distributions based on the total retrievals or a portion of the retrievals would all be candidates. However, given the nature of research, in which often only a modest fraction of projects will achieve their initial objectives, it is crucial to identify those projects that generated a substantial payoff. This suggests emphasis on the top layer of performing projects. This layer could be a fixed number (e.g., top ten) or a percentage of the total (e.g., top 1%). The Finland study we conducted last year [59] used both for comparison purposes, and the relative standings remained the same.

Thus, the citation performance of the ten most cited papers for each technology for each country was compared. Initially, both the median citations and the citations of the two highest papers were used as metrics to obtain multiple perspectives for comparison. However, in many cases the most cited paper was an outlier, and included authors from other (more technologically advanced) countries (especially in India's case). Since the contribution of the authors from other countries to the quality of the target paper was unknown, it was believed that giving full weight to the outliers' citations for either India or China would distort the results. All the top ten papers were retained for computing the median, reflecting the reality that India or China did play some role in the outliers' quality, and the median of the top ten was the final metric employed.

3.3.4. Citation trends

Citations tend to be a widely used, albeit imperfect, proxy metric for quality. In order to track the quality of each country's research output over time, trends in aggregate country citations were followed. Since no single citation metric was adequate as a stand-alone representation of quality, a suite of metrics was employed to provide multiple perspectives on quality.

For every fifth year starting in 1980, the numbers of papers published for each country were tabulated. Then, the numbers of papers with more than a hundred cites were listed, followed by the median of the top twenty cited papers, the median of the top 1% cited papers, and the median citations of the total country output. This approach also allowed the trends in fractions of total publications that were highly cited to be tracked over time, to estimate whether growth in sheer numbers of publications is being achieved at the expense of quality.

3.4. Computational Linguistics

3.4.1. Research Investment Strategy

Understanding India's or China's research investment allocations is the first step to understanding their research investment strategy. This understanding can provide the context for the specific technical thrusts identified. Two approaches were used to determine India's or China's research investment allocations. Both derive from computational linguistics, with the first based on document groupings, and the second based on phrases.

3.4.2. Taxonomy using Document Clustering

The first approach uses document clustering to generate technical categories and the numbers of documents that populate these categories. This approach does not use funding allocations directly, but rather the proxy metric of published documents as a reflection of levels of effort.

Document clustering is the grouping of similar documents into thematic categories. Different approaches have been described ([60–69]]). The approach used for the India and China studies in this Special Section is based on a partitional clustering algorithm ([69,70]) contained within a software package named CLUTO. Most of CLUTO's clustering algorithms treat the clustering problem as an optimization process that seeks to maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. CLUTO uses a randomized incremental optimization algorithm that is greedy in nature and has low computational requirements.

In partitional clustering, the number of clusters desired is input, and all documents in the database are included in those clusters. Clustering was conducted for the 2004 to 2005 documents retrieved from the

SCI/SSCI for China, and for the 2005 documents retrieved from the SCI/SSCI for India. There were 256 clusters run for the retrieved articles from China, and these clusters are listed in detail in [25] in the order in which they appear on the hierarchical tree. They were aggregated into a hierarchical taxonomy using a tree generated by the CLUTO software. There were 256 clusters run for the 1991 and 2002 articles for India, listed in [71]. There were 16 clusters run for the retrieved 2005 articles in the India study in this section of TFSC.

In addition, the India study in this section used an upgraded version of the clustering algorithm that included metrics at each node in the hierarchical taxonomy. This capability allows the technical infrastructure to be coupled to the technical themes at every level of aggregation, and is useful to a wide variety of audiences.

3.4.3. Factor analysis

A factor matrix of phrases/words identifies the main technical themes in a database. Those relatively few words/phrases with the highest absolute values of loadings for a given factor constitute the elements of the technical theme for the factor. It is the responsibility of the human analyst to reconstruct the factor theme from the component words or phrases, usually a straightforward procedure.

3.4.4. Relative investment strategy

The second approach for determining China's or India's research investment allocations uses specific phrases derived from the abstract. To place these results in context, the records associated with very specific Chinese or Indian efforts are compared with those of the USA.

The present approach is based on the philosophy that very specific sub-technology areas should be compared to identify precisely where different countries have concentrated their investment. The critical sub-technologies emphasized by each country become the "dots to be connected" for understanding the overall country research investment strategy.

To determine desired specificities of the technology areas, we follow the chain of disaggregation, starting from the top. At the highest level are the research articles for all of China or India. One could compare the number of research articles in a given year with that of, say, the USA, and draw very general conclusions about overall research output. This was essentially the approach of King, in comparing the research output from thirty-one different countries [72]. Very limited information can be obtained from this level of resolution.

At the next level would be research articles for each technology area for a country. Making comparisons at this level for critical technologies provides a much more strategically important view of each country's capabilities [73]. Recent text-mining studies on nanotechnology [28] and energetic materials [unpublished] show that China is advancing rapidly in its research article production in these two critical technologies, and is second only to the USA in research article production. However, even these results aggregated at the critical technology level may be too general for critical investment strategy emphasis analyses. If China is second to the USA, for example, in nanotechnology overall, might there be sub-areas of nanotechnology (e.g., nanocomposites or nanorods), in which China is actually leading the USA in research article production? And what would be the strategic implications of China heavily emphasizing research investment in such specific areas?

Thus, at the next level would be sub-critical technology areas, such as nanocomposites or nanorods in the example above. Further levels of disaggregation are possible, such as "metal nanocomposites" or "heavy metal nanocomposites." The terminal level of resolution used for the comparison depends on the objectives of the study, and the numbers of articles available at the different levels.

This latter approach was used to compare the relative investment allocations of China and India and the USA for the present section, with a resolution at about the critical sub-technology level. Ten thousand articles each from the USA and China (or India) were downloaded from the SCI/SSCI for 2006 (specifically, the 10,000 most recent articles since 30 August 2006). At the time of the download, the total number of USA articles was tabulated and the total number of China (or India) articles was tabulated. Thus, the ratio of USA to China or India articles could be computed.

A phrase frequency analysis was performed on each download, and the phrases were then combined. The ratio of frequencies for each phrase was tabulated. Phrases were ordered by the ratio of occurrence in each country's download. Two bands were considered: phrases that had a large China (or India)/USA frequency ratio and phrases that had a large USA/China (or India) frequency ratio (the opposite ends of the spectrum). Select phrases in these bands were inserted into the SCI/SSCI, and the actual numbers of records that contained these phrases (for the first 8 months of 2006) were obtained.

Because China's overall SCI/SSCI records were about one fourth those of the USA, and India's overall SCI/SSCI records were about one tenth those of the USA, any technical thrust areas (represented by the technical phrases) in which India or China had greater numbers of records in total than the USA reflected a high relative investment in that specific area by India or China. As will be seen in the specific India–China papers in this Special Section, the results in this section are striking, and dramatically illustrate the differences in India's, China's, and the USA's research investment policies.

References

[1] S. Bhattacharya, P. Nath, Using patent statistics as a measure of technological assertiveness: A China–India comparison, Curr. Sci. 83 (1) (2002) 23–29.

[2] World Investment Report 2005: Transnational corporations and the internationalization of R&D. UNCTAD (United Nations). 2005.

[3] World Investment Prospects 2004: The revival of globalization. The Economist Intelligence Unit, The Economist, 2004.

[4] C.K. Prahalad, G. Hamel, The core competence of the corporation, Harvard Bus. Rev. 68 (3) (1990) 79–91.

[5] D.C. Galunic, S. Rodan, Resource recombinations in the firm: knowledge structures and the potential for Schumpeterian innovation, Strateg. Manage. J. 19 (12) (1998) 1193–1201.

[6] R.N. Kostoff, Text mining for global technology watch, in: M. Drake (Ed.), Second edition, Encyclopedia of Library and Information Science, vol. 4, Marcel Dekker, Inc., New York, NY, 2003, pp. 2789–2799.

[7] C.W. Bostian, W.T. Brandon, A.U. Mac Rae, C.E. Mahle, S.A. Townes, Key technology trends — satellite systems, Space Commun. 16 (2–3) (2000) 97–124.

[8] P. Stares, United-States and Soviet military space programs — A comparative-assessment, Daedalus 114 (2) (1985) 127–145.

[9] R.C.W. Hutubessy, P. Hanvoravongchai, T.T.T. Edejer, Diffusion and utilization of magnetic resonance imaging in Asia, Int. J. Technol. Assess. Health Care 18 (3) (2002) 690–704.

[10] B. Mooney, R. Seymour, WTEC panels survey Russian maritime technologies, Mar. Technol. Soc. J. 30 (1) (1996) 71–72.

[11] L.V. McIntire, WTEC panel report on tissue engineering (Reprinted), Tissue Eng. 9 (1) (2003) 3–7.

[12] R. Campbell, H.D. Balzer, J. Berliner, R. Dobson, P. Gregory, Soviet Science and Technology, Foreign Applied Sciences Assessment Center, October 15 1985.

[13] A. Klinger, editor, A. Klinger et. al. Soviet Image Pattern Recognition Research. Jan. 1990. Foreign Applied Sciences Assessment Center, Science Applications International Corp., 10260 Campus Point Drive, San Diego, CA 92121, and 1710 Goodridge Drive, McLean VA 22102.

[14] E.M. Gray, (Ed.), M. Cohn, L.W. Craver, A. Gersho, T. Lookabaugh, F. Pollara, M. Vetterli. Non-US data compression and coding research. November 1993. A Foreign Applied Sciences Assessment Center (FASAC) report prepared for Science Applications International Corporation (SAIC) under U.S. Government sponsorship.

[15] L.J. Lanzerotti, R.C. Henry, H.P. Klein, H. Masursky, G.A. Paulikas, F.L. Scarf, G.A. Soffen, Y. Terzian, Soviet space science research, FASAC Technical Assessment Report FASAC-TAR-3060. Foreign Applied Sciences Assessment Center, 1986.

[16] L.M. Duncan, F.T. Djuth, J.A. Fejer, N.C. Gerson, T. Hagfors, D.B. Newman Jr., R.L. Showen, Soviet ionospheric modification research, Foreign Applied Sciences Assessment Center, Technical Assessment Report, vol. 4040, 1988.

[17] W.J. Spencer, J.Y. Chen, A. Chiang, W. Frieman, E.S. Kuh, J.L. Moll, R.F. Pease, K.C. Saraswat, Chinese microelectronics, Foreign Applied Sciences Assessment Center Technical Assessment Report. Science Applications International Corporation, April 1989.

[18] R.C. Davidson, M.A. Abdou, L.A. Berry, C.W. Horton, J.F. Lyon, P.H. Rutherford, Japanese magnetic confinement fusion research, Foreign Applied Sciences Assessment Center Technical Assessment Report. Science Applications International Corporation, 1990.

[19] P. Zhou, L. Leydesdorff, The emergence of China as a leading nation in science, Res. Policy 35 (1) (2006) 83–104.

[20] L. Leydesdorff, P. Zhou, Are the contributions of China and Korea upsetting the world system of science? Scientometrics 63 (3) (2005) 617–630.

[21] R.N. Kostoff, R. Tshiteya, K.M. Pfeil, J.A. Humenik, G. Karypis, Power source roadmaps using Database Tomography and bibliometrics, Energy 30 (5) (2005) 709–730.

[22] R.N. Kostoff, H.J. Eberhart, D.R. Toothman, Database Tomography for technical intelligence: comparative analysis of the research impact assessment literature and the Journal of the American Chemical Society, Scientometrics 40 (1) (1997) 103–138.

[23] R.N. Kostoff, R.A. DeMarco, Science and technology text mining, Anal. Chem. 73 (13) (2001) 370A–378A.

[24] R.N. Kostoff, T. Braun, A. Schubert, D.R. Toothman, J.A. Humenik, Fullerene roadmaps using bibliometrics and Database Tomography, J. Chem. Inf. Comput. Sci. 40 (1) (2000) 19–39.

[25] R.N. Kostoff, M.B. Briggs, R.L. Rushenberg, C.A. Bowles, M. Pecht, The structure and infrastructure of Chinese science and technology, DTIC Technical Report Number ADA443315, Defense Technical Information Center, Fort Belvoir, VA, 2006, http://www.dtic.mil/.

[26] R.N. Kostoff, J.A. Del Rio, E.O. García, A.M. Ramírez, J.A. Humenik, Citation mining: Integrating text mining and bibliometrics for research user profiling, JASIST 52 (13) (2001) 1148–1156.

[27] R.N. Kostoff, M.F. Shlesinger, G. Malpohl, Fractals roadmaps using bibliometrics and database tomography, Fractals 12 (1) (2004) 1–16.

[28] R.N. Kostoff, J.A. Stump, D. Johnson, J. Murday, C.G.Y. Lau, W. Tolles, The structure and infrastructure of the global nanotechnology literature, J. Nanopart. Res. 8 (3–4) (2006) 301–321.

[29] R.N. Kostoff, H.J. Eberhart, D.R. Toothman, Database tomography for information retrieval, J. Inf. Sci. 23 (4) (1997) 301–311.

[30] E. Greengrass, Information retrieval: An overview, National Security Agency. TR-R52-02-96, 1997.

[31] TREC (Text Retrieval Conference), (2005). Home Page, http://trec.nist.gov/.

[32] D.R. Swanson, Fish Oil, Raynauds Syndrome, and undiscovered public knowledge, Perspect. Biol. Med. 30 (1) (1986) 7–18.

[33] D.R. Swanson, N.R. Smalheiser, An interactive system for finding complementary literatures: a stimulus to scientific discovery, Artif. Intell. 91 (2) (1997) 183–203.

[34] R.N. Kostoff, Stimulating innovation, in: Larisa V. Shavinina (Ed.), International Handbook of Innovation, Elsevier Social and Behavioral Sciences, Oxford, UK, 2003, pp. 388–400.

[35] J.A. Goldman, W.W. Chu, D.S. Parker, R.M. Goldman, Term domain distribution analysis: a data mining tool for text databases, Methods Inf. Med. 38 (1999) 96–101.

[36] R.N. Kostoff, Bilateral asymmetry prediction, Med. Hypotheses 61 (2) (2003) 265–266.

[37] F. Narin, Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity (monograph). NSF C 637. National Science Foundation. Contract NSF C 627. NTIS Accession No. PB252339/AS. 1976.

[38] E. Garfield, History of citation indexes for chemistry — a brief review, JCICS 25 (3) (1985) 170–174.

[39] F. Narin, D. Olivastro, K.A. Stevens, Bibliometrics theory, practice and problems, Eval. Rev. 18 (1) (1994) 65–76.

[40] Z. Jiang, Hold high the great banner of Deng Xiaoping theory for an all-round advancement of the cause of building socialism with Chinese characteristics into the 21st Century, Report Delivered at the 15th National Congress of the Communist Party of China, September 12 1997.

[41] People's Daily Online, Chinese President on development of science and technology, People's Daily Online, 18 June 2000, http://english.peopledaily.com.cn/.

[42] Chinese Embassy, Science and Technology Policy, 2005.

[43] Cox Report, House Report 105–851. Report of the Select Committee on U.S. national security and military/commercial concerns with the People's Republic of China. Rep.Christopher Cox of California, Chairman, United States Congress. 14 June 1999, http://www.access.gpo.gov/congress /house/hr105851.

[44] MOST, Science and Technology Indicators: 2002, Scientific and Technical Documents Publishing House, Beijing, 2005.

[45] MOST, China science and Technology Statistics Data Book, http://www.most.gov.cn/eng/ statistics/2003/index.html, 2003.

[46] D.L. Hsiung, An Evaluation of China's Science and Technology System and its Impact on the Research Community, 2002.

[47] W. Blanpied (Ed.), Proceedings of the Sino-US Forum on Basic Research for the Next Fifteen Years, 2002.

[48] Xinhua, China rises to third in research, development spending, http://www1.chinadaily.com.cn/en/doc/2003-11/03/content_277967.htm, November 3 2003.

[49] G.H. Jefferson, Impact of the Foreign Sector on Innovation in China's Domestic Firm, 2006.

[50] National Science Board (NSB), Science and Engineering Indicators, 2004.

[51] Research and Development Statistics 2004–05. Department of Science and Technology. Government of India, Report by Comptroller and Auditor General (2002), Report No. 5, Government of India, 2006.

[52] SCI, Certain data included herein are derived from the Science Citation Index/Social Science Citation Index prepared by the THOMSON SCIENTIFIC ®, Inc. (Thomson®), Philadelphia, Pennsylvania, USA: © Copyright THOMSON SCIENTIFIC ® 2006. All rights reserved.

[53] EC, Compendex®, Engineering Village, Elsevier Engineering Information, Inc. Hoboken, NJ 07030, 2006.

[54] R.B. Cattell, The Screen Test for the number of factors, Multivariate Behav. Res. 1 (1966) 245–276.

[55] R.N. Kostoff, J.A. Block, Factor matrix text filtering and clustering, JASIST 56 (9) (2005) 946–968.

[56] H.F. Kaiser, The application of electronic computers to factor analysis, Educ. Psychol. Meas. 20 (1960) 141–151.

[57] R.N. Kostoff, The use and misuse of citation analysis in research evaluation, Scientometrics 43 (1) (1998) 27–43.

[58] R.N. Kostoff, The handbook of research impact assessment, DTIC Technical Report Number ADA296021, Seventh edition, Summer 1997.

[59] R.N. Kostoff, R. Tshiteya, C.A. Bowles, T. Tuunanen, The structure and infrastructure of the Finnish research literature, Technol. Anal. Strateg. Manag. 18 (2) (2006) 187–220.

[60] D.R. Cutting, D.R. Karger, J.O. Pedersen, J.W. Tukey, Scatter/Gather: A cluster-based approach to browsing large document collections, Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92), 1992, pp. 318–329.

[61] S. Guha, R. Rastogi, K. Shim, CURE: An efficient clustering algorithm for large databases, Proceedings of the ACM-SIGMOD 1998 International Conference on Management of Data (SIGMOD'98), 1998, pp. 73–84.

[62] M.A. Hearst, The use of categories and clusters in information access interfaces, in: T. Strzalkowski (Ed.), Natural Language Information Retrieval, Kluwer Academic Publishers, 2000.

[63] G. Karypis, E.H. Han, V. Kumar, Chameleon: A hierarchical clustering algorithm using dynamic modeling, IEEE Comput. Special Issue Data Anal. Min. 32 (8) (1999) 68–75.

[64] E. Rasmussen, Clustering algorithms, in: W.B. Frakes, R. Baeza-Yates (Eds.), Information Retrieval Data Structures and Algorithms, Prentice Hall, N.J., 1992.

[65] M. Steinbach, G. Karypis, V. Kumar, A comparison of document clustering techniques, Technical Report #00-034, 2000. Department of Computer Science and Engineering. University of Minnesota.

[66] P. Willet, Recent trends in hierarchical document clustering: a critical review, Inf. Process. Manag. 24 (1988) 577–597.

[67] O. Zamir, O. Etzioni, Web document clustering: A feasibility demonstration, Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), 1998, pp. 46–54.

[68] Y. Zhao, G. Karypis, Empirical and theoretical comparisons of selected criterion functions for document clustering, Mach. Learn. 55 (3) (2004) 311–331.

[69] Y. Zhao, G. Karypis, Hierarchical clustering algorithms for document datasets, Data Min. Knowl. Discov. 10 (2) (2005) 141–168.

[70] G. Karypis, CLUTO—A clustering toolkit, http://www.cs.umn.edu/~cluto, 2005.

[71] R.N. Kostoff, D. Johnson, C.A. Bowles, S. Dodbele, Assessment of India's research literature, DTIC Technical Report ADA444625, Defense Technical Information Center, Fort Belvoir, VA, 2006, http://www.dtic.mil/.

[72] D.A. King, The scientific impact of nations, Nature 430 (6997) (2004) 311–316.

[73] R.N. Kostoff, Scientific impact of nations, The Scientist (September 27 2004).

Ronald N. Kostoff received a Ph.D. in Aerospace and Mechanical Sciences from Princeton University in 1967. He has worked for Bell Laboratories, Department of Energy, and Office of Naval Research (ONR). He has authored over 100 technical papers, served as Guest Editor of three journal Special Issues, obtained two text mining system patents, and presently manages a text mining pilot program at ONR.

Sujit Bhattacharya received a Ph.D. from the Indian Institute of Technology in Informatics, and presently is a scientist in the National Institute of Science Tecnhnology and Development Studies (NISTADS). His main area of research is in Scientometrics and Informetrics (Patent-Related Studies), and Technology, Trade and IPR issues. He has published extensively, including two authored books and one co-edited book.

Michael Pecht has an M.S. and Ph.D. in Engineering Mechanics from the University of Wisconsin at Madison. He is a Professional Engineer, an IEEE Fellow, and an ASME Fellow. In addition to writing eighteen books on electronic products development, use and supply chain management, he has received the 3M Research Award for electronics packaging, the IEEE Undergraduate Teaching Award, and the IMAPS William D. Ashman Memorial Achievement Award for his contributions in electronics reliability analysis.