Available online at www.sciencedirect.com
ScienceDirect
Technological Forecasting & Social Change 74 (2007) 1574–1608
Technological Forecasting and Social Change

Assessment of India's research literature ☆

Ronald N. Kostoff [a,*], Dustin Johnson [a,1], Christine A. Bowles [b], Sujit Bhattacharya [c], Alan S. Icenhour [a], Kimberly Nikodym [a], Ryan B. Barth [b], Simha Dodbele [a]

[a] *Office of Naval Research, 875 N. Randolph St., Arlington, VA 22217, USA*
[b] *DDL-OMNI Engineering, LLC, 8260 Greensboro Drive, Suite 600, Mclean, VA 22102, USA*
[c] *National Institute of Science, Technology and Development Studies (NISTADS), Pusa Gate, K.S. Krishnan Marg, New Delhi-110012, India*

Received 28 October 2006; received in revised form 13 February 2007; accepted 16 February 2007

Abstract

The structure and infrastructure of the Indian research literature were determined. A representative database of technical articles was extracted from the Science Citation Index/Social Science Citation Index (SCI/SSCI) [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. [1]] for 2005, with each article containing at least one author with an India address. Document clustering was used to identify the main technical themes (core competencies) of Indian research. Aggregate India bibliometrics were also performed, emphasizing the value of collaborative research to India. A unique mapping approach was used to identify networks of organizations that published together, networks of organizations with common technical interests, and especially those organizations with common technical interests that did not co-publish extensively. Finally, trend analyses were performed using other year data from the SCI/SSCI to place the 2005 results in their proper historical context. Published by Elsevier Inc.

Keywords: India; Science and technology; Technology assessment; Core competencies; Research evaluation; Metrics; Bibliometrics; Text mining; Computational linguistics; Document clustering; CLUTO; Auto-correlation mapping; Cross-correlation mapping; Factor analysis; Factor matrix; Impact Factor

☆ The views in this paper are solely those of the authors, and do not necessarily represent the views of the Department of the Navy or any of its components, DDL-OMNI Engineering, LLC, Northrop Grumman, or the National Institute of Science, Technology and Development Studies.

* Corresponding author. Tel.: +1 703 696 4198; fax: +1 703 696 8744.
E-mail address: kostofr@onr.navy.mil (R.N. Kostoff).
[1] Presently, Northrop Grumman TASC, 12015 Lee Jackson Highway, Fairfax, VA 22033, United States.

0040-1625/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.techfore.2007.02.009

1. Introduction

South–East and East Asia have become dynamic growth areas, especially in science and technology (S&T) (see for example [2]). Our text mining studies of specific technologies over recent years have shown dramatic growth in research output production by China, South Korea, Taiwan, and Singapore (e.g., [3]), to name a few. As a result, we have started to adopt a national view of research output from some countries in the region, and are examining research products from individual countries. The preceding paper in this Special Issue was focused on an assessment of China's research enterprise. The present paper focuses on India's research enterprise, and the next paper in this Special Issue will compare the research outputs of the two countries.

The primary objective of the present study is to identify the S&T core competencies of India. In addition, temporal trends of significant research-related parameters will be presented. These trends will provide a context in which to interpret India's present research output status, and will provide support for the predictive conclusions that follow.

2. Background

The present study combines three concepts/approaches for the assessment of India's S&T literature: core competency determination, country technology assessments, and text mining assessments. The background for these three concepts, as well as a description of India's S&T enterprise, can be found in the Introduction of this Special Issue. India's S&T performance based on research output literature will now be summarized.

3. Approach and results

3.1. Overview

A taxonomy and detailed bibliometrics analyses are presented for 1 year (2005). Gross bibliometric trends are presented to place the detailed 2005 bibliometrics in perspective. The databases used for the bibliometrics and taxonomy analyses, the bibliometrics approaches, and the document clustering taxonomy approach, are described in the Introduction of this Special Issue.

3.2. Bibliometrics

Publication and citation bibliometrics were performed at the aggregate national level. In addition, bibliometrics of four core technologies were examined, and can be found in detail in [4].

3.2.1. Overall India bibliometrics

This section presents temporal publication trends, journals containing most articles, journals cited most frequently by Indian authors, most prolific institutions, and most collaborative countries for the aggregate India database.

3.2.1.1. Publication trends. The first metric is number of articles as a function of time. All research articles in the SCI/SSCI having at least one author with an India address were retrieved for selected years

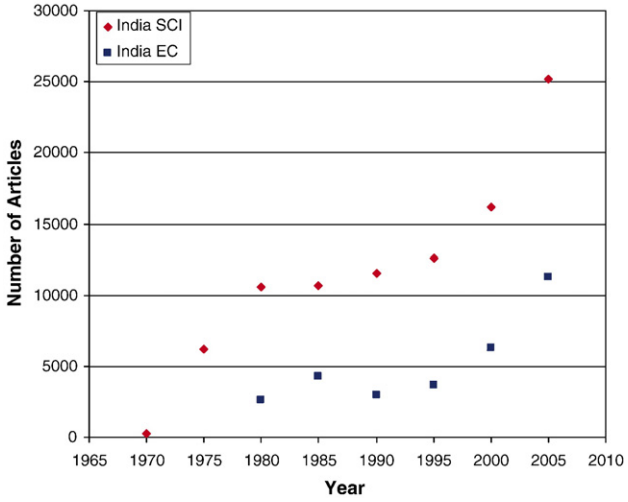

Fig. 1. Number of SCI and EC Research Articles with Indian Authors.

and the results are shown in Fig. 1. In addition, all articles in the Engineering Compendex (EC) [5] having an India address in the author affiliation field were retrieved for selected years. The publishing rate reached a plateau during the period 1980 to 1995, and it has since started a rapid increase.

The SCI output contains a field called Subject Category. It is essentially a classification by technical thrust of each article. Table 1 lists the top ten Subject Categories for all of the Indian research articles since 1980. In 1980, a broad number of topics are represented, including chemistry, physics, plant science, and medical-related topics. By 2005, the topics have become much more focused—primarily in the areas of chemistry, physics, and materials.

The EC output contains the thematic fields Classification Codes and Controlled Vocabulary. Table 2 is the temporal trend of Classification Codes. While there has been a strong effort in chemical-related topics throughout the time period shown, this chemical focus has strengthened more in recent times. Physics and mathematics appear to constitute the second tier. Where are information technology, and mechanics and properties of materials?

Table 3 is the temporal trend of Controlled Vocabulary. The perspective provided on Indian technology is very different from that provided by Table 2, at least superficially. The recent focus appears concentrated on applied mathematics (mathematical models, computer simulation, algorithms, optimization) and nanotechnology-focused technologies (X-ray diffraction analysis, nanostructured materials, scanning electron microscopy, thin films). In order for congruence between the two perspectives to exist, the nanotechnology topics are probably being classified under the chemical-related codes. Additionally, the applied mathematics must be providing substantial support to solving chemical-related problems. These two perspectives show the importance of employing multiple metrics when assessing research.

3.2.1.2. Journals

3.2.1.2.1. Journals containing most Indian-authored articles. The journals containing the most research articles with at least one Indian author (from the total 2005 database using the SCI Analyze function) are shown on Table 4.

Table 1
SCI Subject Categories as a Function of Time

1980		1985		1990		1995		2000		2005	
Chemistry, Multidisciplinary	903	Chemistry, Multidisciplinary	650	Chemistry, Multidisciplinary	627	Materials Science, Multidisciplinary	876	Materials Science, Multidisciplinary	960	Materials Science, Multidisciplinary	1634
Physics, Multidisciplinary	625	Physics, Multidisciplinary	538	Materials Science, Multidisciplinary	563	Chemistry, Multidisciplinary	692	Chemistry, Multidisciplinary	904	Chemistry, Multidisciplinary	1553
Plant Sciences	546	Plant Sciences	510	Physics, Multidisciplinary	477	Chemistry, Physical	587	Chemistry, Organic	806	Chemistry, Organic	1542
Multidisciplinary Sciences	480	Biochemistry & Molecular Biology	432	Physics, Condensed Matter	472	Physics, Condensed Matter	579	Chemistry, Physical	731	Chemistry, Physical	1470
Medicine, General & Internal	451	Multidisciplinary Sciences	408	Chemistry, Physical	458	Biochemistry & Molecular Biology	524	Physics, Multidisciplinary	682	Biochemistry & Molecular Biology	1166
Biochemistry & Molecular Biology	392	Chemistry, Organic	367	Chemistry, Organic	453	Physics, Multidisciplinary	501	Biochemistry & Molecular Biology	666	Physics, Condensed Matter	971
Immunology	381	Agronomy	346	Biochemistry & Molecular Biology	448	Chemistry, Organic	463	Physics, Condensed Matter	549	Physics, Multidisciplinary	953
Agronomy	375	Chemistry, Inorganic & Nuclear	316	Plant Sciences	424	Physics, Applied	457	Multidisciplinary Sciences	537	Physics, Applied	802
Chemistry, Organic	374	Physics, Condensed Matter	312	Physics, Applied	382	Engineering, Chemical	385	Agriculture, Dairy & Animal Science	503	Engineering, Chemical	788

Table 2
EC Classification Codes as a Function of Time

Engineering Compendex classification code for Indian journal articles									
1985		1990		1995		2000		2005	
Chemical Products Generally	1019	Chemical Products Generally	837	Numerical Methods	902	Chemical Reactions	943	Chemical Reactions	2719
Applied Physics Generally	675	Applied Mathematics	511	Inorganic Compounds	631	Inorganic Compounds	923	Organic Compounds	2505
Chemical Apparatus and Plants; Unit Operations; Unit Processes	663	Applied Physics Generally	490	Physical Properties of Gases, Liquids & Solids	594	Physical Properties of Gases, Liquids & Solids	790	Chemical Operations	2155
Applied Mathematics	622	Chemical Apparatus and Plants; Unit Operations; Unit Processes	431	Chemical Reactions	543	Organic Compounds	721	Inorganic Compounds	2154
Electricity and Magnetism	412	Electricity and Magnetism	387	Electricity: Basic Concepts & Phenomena	502	Atomic & Molecular Physics	702	Physical Properties of Gases, Liquids & Solids	1816
Chemistry	384	Computer Software, Data Handling and Applications	309	Chemical Operations	436	Chemical Operations	689	Atomic & Molecular Physics	1654
Strength of Building Materials; Mechanical Properties	368	Light, Optics and Optical Devices	297	Physical Chemistry	428	Numerical Methods	687	Light/Optics	1595
Computer Software, Data Handling and Applications	321	Chemistry	277	Strength of Building Materials; Mechanical Properties	378	Physical Chemistry	610	Electricity: Basic Concepts & Phenomena	1500
Metallurgy and Metallography	278	Metallurgy and Metallography	242	Computer Applications	371	Applied Mathematics	579	Applied Mathematics	1491
Light, Optics and Optical Devices	277	Strength of Building Materials; Mechanical Properties	238	Mechanics	337	Electricity: Basic Concepts & Phenomena	557	Physical Chemistry	1479

Table 3
EC Controlled Vocabulary as a Function of Time

Engineering Compendex controlled vocabulary for Indian journal articles										
1985		1990		1995		2000		2005		
Mathematical Models	78	Mathematical Models	81	Mathematical Models	655	Mathematical Models	679	Mathematical Models	1278
Chemical Reactions — Reaction Kinetics	47	Ceramic Materials	54	Algorithms	249	Computer Simulation	416	Synthesis Chemical	906
Mathematical Techniques	47	Mathematical Techniques—Finite Element Method	40	Thermal Effects	241	Thermal Effects	344	Computer Simulation	796
Electrochemistry	46	Spectroscopy, Infrared	39	Computer Simulation	240	Synthesis Chemical	306	X Ray Diffraction Analysis	553
Mathematical Techniques — Finite Element Method	41	Computer Simulation	35	Performance	171	Algorithms	258	Nanostructured Materials	533
Computer Programming — Algorithms	39	Computer Programming—Algorithms	32	Calculations	159	Composition Effects	231	Algorithms	524
Thermodynamics	35	Oxides	32	Synthesis Chemical	155	X Ray Diffraction Analysis	217	Optimization	445
Stresses — Analysis	31	Thermal Effects	31	X ray Diffraction	154	Scanning Electron Microscopy	160	Scanning Electron Microscopy	442
Crystals — Structure	30	Chemical Reactions—Reaction Kinetics	31	Scanning Electron Microscopy	151	Reaction Kinetics	153	Reaction Kinetics	432
Probability	30	X-ray Analysis	28	Composition	151	Thin Films	148	Thin Films	382

The highest ranking journals (*ranked in terms of frequency of papers containing at least one Indian author*) emphasize veterinary and animal science, chemistry, agriculture, physics, and materials, in that order. The strong material emphasis shown in Table 1 is not fully evident in Table 4. For compatibility, this means that much of the chemistry research must be materials-related (compatible with the materials-oriented nanotechnology focus discussed in the Controlled Vocabulary section) and/or there are substantial numbers of specialty materials journals with low frequencies in the total journal lists. Fifteen of the 25 journals listed are domestic Indian journals. The journal Impact Factors are relatively low; seventeen of the 25 journals listed have Impact Factor less than unity. The median date when the 25 journals were accessed initially by the SCI/SSCI was 1973.

The weighted Impact Factor is calculated by evaluating the Impact Factor for the top 10 journals for a given year and then calculating an average that is weighted by the number of publications in each journal. The variation with time in the weighted Impact Factor for journals containing Indian articles is shown in Fig. 2. The circles reflect the computation of weighted Impact Factors that omit journals with no reported Impact Factor, while the squares include the journals that have no reported Impact Factor (i.e., the Impact Factor for those journals is taken as zero). The weighted Impact Factor is relatively unchanged (or shows a slight decrease) during the period 1980 to 2000. The large increase in 2005 reflects increased publication in higher quality journals.

3.2.1.2.2. High Impact Factor journals. How does collaboration among India and other countries impact the journals in which Indian authors publish? A very brief analysis was performed. Two cases were

Table 4
Journals Containing Most Articles by Indian authors (Retrieved 2005 database)

Indian journal	# PAP	IMP FACT	THEME	SCI ACC DATE
Current Science	457	0.728	Multidisciplinary	1961
Indian Veterinary Journal	443	0.052	Vet	1977
Indian Journal Of Animal Sciences	381	0.09	Vet	1976
Asian Journal Of Chemistry	346	0.153	Chem	1995
Tetrahedron Letters	272	2.477	Chem	1959
Journal Of The Indian Chemical Society	267	0.34	Chem	1946
Acta Crystallographica Section E—Structure Reports Online	242	0.581	Matls	2001
Indian Journal Of Chemistry Section B—Organic Chemistry Including Medicinal Chemistry	240	0.446	Chem	1976
Journal Of Food Science And Technology—Mysore	217	0.123	Agri	1976
Physical Review B	187	3.185	Physics	1964
Indian Journal Of Agricultural Sciences	172	0.084	Agri	1966
Indian Journal Of Physics And Proceedings Of The Indian Association For The Cultivation Of Science	170	0.072	Physics	1968
Pramana-Journal Of Physics	146	0.38	Physics	1990
Indian Journal Of Chemistry Section A—Inorganic Bio-Inorganic Physical Theoretical & Analytical Chemistry	138	0.632	Chem	1976
Indian Journal Of Pure & Applied Physics	134	0.495	Physics	1964
Journal Of Applied Polymer Science	134	1.072	Matls	1965
Journal Of Applied Physics	132	2.498	Physics	1937
Spectrochimica Acta Part A—Molecular And Biomolecular Spectroscopy	122	1.29	Chem	1973
Journal Of The Geological Society Of India	115	0.217	Geol	1970
Indian Journal Of Heterocyclic Chemistry	114	0.312	Chem	1995
Bulletin Of Materials Science	109	0.777	Matls	1986
Physical Review D	109	4.852	Physics	1970
Journal Of Physical Chemistry B	107	4.033	Chem	1997
Physica B—Condensed Matter	107	0.796	Physics	1990
Physical Review Letters	105	7.489	Physics	1958

examined and compared. The first case represents articles that could have included participation among India and other countries. The second case represents articles published essentially exclusively by Indian authors. The differences between the two cases represent the impact of collaboration.

In the first case, all research articles in the SCI/SSCI having at least one author with an India address, and publication date of 2005, were retrieved. There were 25,367 records. In the second case, all research articles in the SCI/SSCI having at least one author with an India address, a publication date of 2005, but excluding authors from India's 25 major collaborators, were retrieved. There were 20,672 records retrieved, a 20% reduction from the collaborative case. Thus, it can be concluded that collaboration contributed 20% to the overall publication output.

A small sample of high Impact Factor journals was examined. Table 5 lists these journals.

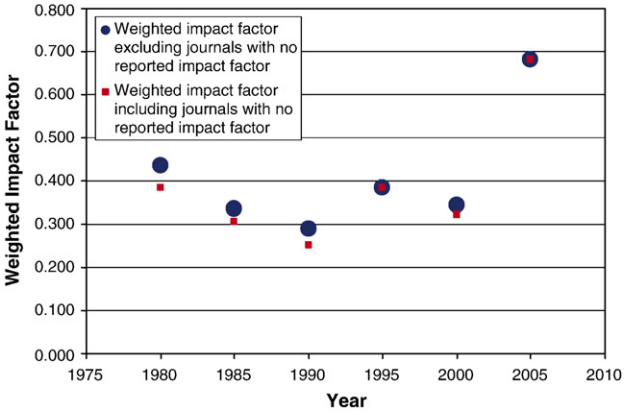

Fig. 2. Weighted Impact Factor for the Top 10 Journals with Indian Authors.

Collaboration has the effect of dramatically increasing the presence of papers with Indian authors in the higher Impact Factor journals. The effect of collaboration on citations will be addressed in the section on collaborative countries. It is equally dramatic!

Additionally, a very short experiment was performed to estimate growth of India articles in high Impact Factor journals. Three of the five most highly cited journals (one in each of the following disciplines: chemistry, physics, biology) was selected, and the numbers of papers published with Indian authors were examined as a function of time. Table 6 contains the results.

Prior to the mid-1990s, the number of India articles in these three journals were relatively low. Therefore, only the data from the mid-1990s to the present are shown. Over the decade, growth in all three journals has been substantial.

How does the growth in these three highly cited journals compare with the overall research article growth of India (shown previously) in this period? From 1995 to 2000, India's overall article growth increased by about a third (16,203/12,602). In Table 6, for all three journals, India's growth over this period is greater than a third, ranging from factors of 1.5 to four. From 2000 to 2005, India's overall article growth was about 50% (25,367/16,203). In Table 1, India's article growth ranged from factors of zero to 2.5, on average matching its overall article growth during this period.

The message to be taken from this analysis is that India is increasing its growth of articles in highly cited journals greater than its overall increase in growth of research articles. India's relative increase is modest, and much of the increase in overall research article growth comes from increasing production of articles in low Impact Factor domestic and international journals. Also, there is increased production

Table 5
High Impact Factor Journals (Total 2005 records)

JOURNAL	INDIA ONLY	INDIA & COLLAB	IMPACT FACTOR
Nature	1	8	32.18
Science	2	8	31.85
Physical Review Letters	25	106	7.22
PNAS-USA	5	14	10.45

Table 6
India Publications in Selected Journals

YEAR	JACS INDIA	P REV LETT INDIA	J BIO CHEM INDIA
1995	5	34	9
1996	17	45	13
1997	16	54	17
1998	23	66	9
1999	13	55	23
2000	19	63	16
2001	15	72	41
2002	15	81	30
2003	18	75	56
2004	14	84	54
2005	29	105	51

CODE:
JACS=Journal of the American Chemical Society.
P REV LETT=Physical Review Letters.
J BIO CHEM=Journal of Biological Chemistry.

in high Impact Factor journals. The increase in high Impact Factor journals outpaces the increase in overall research article production, but the high Impact Factor journal production is a relatively small fraction of the overall research article production.

Table 7
Journals Most Cited by Indian Authors (Retrieved 2005 database)

JOURNAL	# CITES	IMP FACT	THEME
J Am Chem Soc	5559	6.9	CHEM
Phys Rev Lett	4494	7.22	PHYS
Phys Rev B	3835	3.08	PHYS
Nature	3399	32.18	SCIENCE
J Biol Chem	3058	6.36	CHEM
Science	2834	31.86	SCIENCE
Tetrahedron Lett	2809	2.48	CHEM
J Chem Phys	2704	3.11	CHEM
J Org Chem	2541	3.46	CHEM
P Natl Acad SCI USA	2299	10.45	SCIENCE
Phys Rev D	2258	5.16	PHYS
Inorg Chem	2144	3.45	CHEM
J Phys Chem—US	2036	2.81	PHYS
J Appl Phys	1758	2.26	PHYS
Appl Phys Lett	1635	4.31	PHYS
Chem Rev	1558	20.23	CHEM
Angew Chem Int Edit	1544	9.16	CHEM
J Phys Chem B	1465	3.83	PHYS
Tetrahedron	1465	2.64	CHEM
Phys Lett B	1421	4.62	PHYS
J Appl Polym Sci	1417	1.021	MAT'LS

Table 8
Top 25 Indian Institutions Based on Total Number of Publications for 1985, 1995, and 2005

1985		1995		2005	
Institution Name	Record Count (10632)	Institution Name	Record Count (12603)	Institution Name	Record Count (25227)
Indian Inst Technol	1105	Indian Inst Technol	1483	Indian Inst Technol	2986
Indian Inst Sci	427	Indian Inst Sci	649	Indian Inst Sci	1110
Banaras Hindu Univ	345	Bhabha Atom Res Ctr	390	Bhabha Atom Res Ctr	648
Bhabha Atom Res Ctr	310	Banaras Hindu Univ	344	Univ Delhi	507
Univ Delhi	222	Tata Inst Fundamental Res	292	Indian Inst Chem Technol	468
Tata Inst Fundamental Res	206	Univ Delhi	259	Tata Inst Fundamental Res	465
Haryana Agr Univ	179	Indian Assoc Cultivat Sci	208	All India Inst Med Sci	425
Univ Roorkee	177	Natl Chem Lab	208	Natl Chem Lab	420
Punjab Agr Univ	153	Jadavpur Univ	179	Jadavpur Univ	402
Panjab Univ	142	Univ Madras	155	Banaras Hindu Univ	342
Univ Calcutta	140	Univ Calcutta	152	Univ Madras	321
Natl Chem Lab	139	All India Inst Med Sci	146	Indian Assoc Cultivat Sci	309
Aligarh Muslim Univ	136	Indian Stat Inst	138	Anna Univ	301
All India Inst Med Sci	134	Univ Hyderabad	137	Panjab Univ	270
Jadavpur Univ	123	Univ Roorkee	126	Aligarh Muslim Univ	252
Cent Drug Res Inst	122	Aligarh Muslim Univ	124	Indian Stat Inst	244
Postgrad Inst Med Educ & Res	114	CSIR	115	Univ Hyderabad	237
Indian Agr Res Inst	113	Natl Phys Lab	112	CSIR	235
Andhra Univ	104	Osmania Univ	112	Univ Calcutta	232
Indian Assoc Cultivat Sci	103	Saha Inst Nucl Phys	109	Postgrad Inst Med Educ & Res	224
Indian Stat Inst	100	Punjabi Univ	107	Natl Inst Technol	218
Osmania Univ	100	Punjab Agr Univ	104	Cent Drug Res Inst	207
Sri Venkateswara Univ	100	Indian Inst Chem Technol	101	Annamalai Univ	195
Univ Rajasthan	100	Indira Gandhi Ctr Atom Res	99	Univ Mysore	190
Univ Madras	90	Cent Drug Res Inst	98	Saha Inst Nucl Phys	185

3.2.1.2.3. Most cited journals. For the overall country citation metrics, the citations in all the retrieved SCI/SSCI papers were aggregated. The journals cited most frequently were identified, and are presented in Table 7 in order of decreasing frequency.

Two important features stand out. First, all the journals in Table 7 are international journals; none are Indian. Second, the Impact Factors for these most cited journals are almost an order of magnitude higher than the Impact Factors of the journals that contain the most Indian papers. Thus, Indian authors are citing

the high Impact Factor journals extensively, but not publishing in them extensively. As was shown in the previous section, Indian authors are increasing their presence in these high Impact Factor journals, but they are presently over-concentrated in the lower Impact Factor journals.

3.2.1.3. Prolific institutions

3.2.1.3.1. List of most prolific institutions. Table 8 lists the top 25 institutions for 1985, 1995, and 2005, allowing changes in institutional rankings with time to be determined. There are 22 universities/engineering colleges and 13 research institutes in the list of prolific institutions. There is little change in the most prolific institutions. However, several agriculture-related institutions that were top publishers in 1985 are not on the list in 2005, reflecting the trend towards chemistry, physics, and materials. The top institutions display a steady increase in publications, with the top institutions roughly doubling their annual output from 1995 to 2005. Institutions that were not able to achieve this level of growth were either not among the top 25 institutions in 2005 (this happened mainly for institutions that had lower rank in 1985) or were at a much lower rank then they had in 1985.

Two institutions stand out in terms of productivity: Indian Institute of Technology (IIT) and Indian Institute of Science. However, it should be noted that output of IIT is the total aggregate of six IITs. All the six IITs are premier engineering colleges in India with central funding and similar areas of research focus. Another institute i.e. National Institute of Technology (NIT) is visible in the list of prolific institutes. Twenty engineering colleges are under NIT. They are in each major state and, like IITs, receive central funding. They are considered next in ranking after IITs.

The increase in number of publications was examined in terms of the Impact Factor. Weighted Impact Factors were calculated for the top 10 institutions for 1985, 1995, and 2005, as shown in Table 9. (The weighted Impact Factor was computed based on the top 10 journals in which an institution published for a given year.) The table shows that the Impact Factor is increasing for the top institutions—somewhat dramatically in some cases, which is consistent with the data presented in Fig. 2. The increase is significant when one considers the overall increase in publishing rate. Thus, the Indian authors are not only increasing their publication rate, they are doing so in higher quality journals.

3.2.1.3.2. Institution auto-correlation map. How do these institutions collaborate? Fig. 3 is an auto-correlation map of the ~30 most prolific institutions (generated by the TechOasis software). No *strongly*

Table 9
Weighted Impact Factors of Top 10 Indian Research Institutions for Selected Years (Order of Institutions based on the Number of Publications during a Given Year)

1985		1995		2005	
Indian Inst Technol	0.746	Indian Inst Technol	1.177	Indian Inst Technol	1.654
Indian Inst Sci	1.047	Indian Inst Sci	1.469	Indian Inst Sci	2.490
Banaras Hindu Univ	0.635	Bhabha Atom Res Ctr	1.028	Bhabha Atom Res Ctr	2.759
Bhabha Atom Res Ctr	0.536	Banaras Hindu Univ	0.875	Univ Delhi	2.518
Univ Delhi	0.833	Tata Inst Fundamental Res	4.169	Indian Inst Chem Technol	1.604
Tata Inst Fundamental Res	2.456	Univ Delhi	2.590	Tata Inst Fundamental Res	4.925
Haryana Agr Univ	0.348	Indian Assoc Cultivat Sci	1.570	All India Inst Med Sci	1.439
Univ Roorkee	0.682	Natl Chem Lab	1.439	Natl Chem Lab	2.344
Punjab Agr Univ	0.463	Jadavpur Univ	1.428	Jadavpur Univ	1.236
Panjab Univ	0.853	Univ Madras	0.450	Banaras Hindu Univ	2.072

connected publishing groupings or even linkages are evident, but five *moderately connected* publishing groupings can be identified:

- University of Madras—centered group (top center)
- Punjab University—centered group (mid-left)
- University of Calcutta—centered group (bottom center)

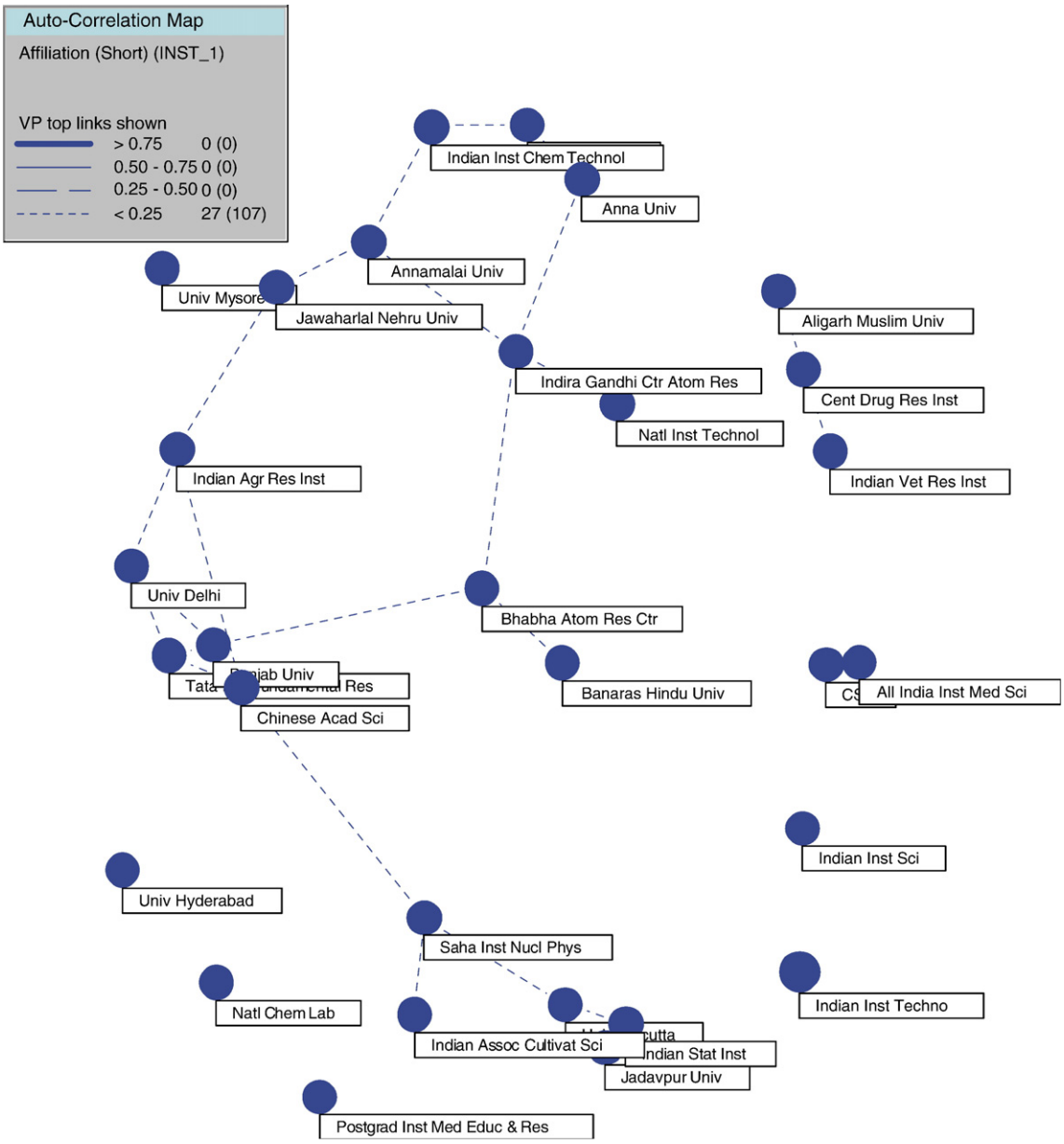

Fig. 3. Institution Auto-Correlation Map (based on most prolific institutions).

- Bhabha ACR-centered group (mid-center)
- Because of its sheer magnitude, the Indian Institute of Technology (actually, an aggregate of six IITs within the country) has to be included as a self-contained group.

In addition to the intra-connection within the first four groups, there is reasonable inter-connection across these four groups evident from this diagram. However, a number of institutions, including the two most prolific producers of research articles, do not show external connections on this specific diagram, given the selected threshold connectivity level for displaying linkages.

It should be emphasized that the auto-correlation map is only one piece in a larger puzzle. The relationships shown are intra-country (with one exception). As research becomes more global, national boundary lines must be crossed to reveal the full extent of inter-country collaboration.

As a very small illustrative example, the Chinese Academy of Sciences was included for mapping purposes. In Fig. 3, it is shown weakly linked to Panjab University and Tata Institute for Fundamental Research. The reasons for this linkage will be discussed in the analysis of the cross-correlation maps.

To display these groupings more quantitatively, a factor analysis was performed on the institutions listed in Table 8.

3.2.1.3.3. Institution factor matrix. A factor analysis was performed to identify the strength of the publishing linkages among institutions. A five factor model was selected, based on the groupings shown in Fig. 3. Five distinct groupings are shown, one for each factor.

- University of Madras strongly linked to the Indian Institute of Chemical Technology, and weakly linked to Anna University, Punjab University, Tata Institute of Fundamental Research (TATA IFR), and Chinese Academy of Sciences,
- Punjab University strongly linked to Tata IFR, and weakly linked to University of Delhi.
- University of Calcutta strongly linked to Jadavpur University, Saha Institute of Nuclear Physics (Shah INP), and Indian Statistical Institute, and weakly linked to Indian Association of Cultivation of Science.
- Indian Institute of Technology with very weak links (same sign of factor loadings) to National Institute of Technology, University of Calcutta, Indira Gandhi CAR, and University of Madras.
- Bhabha Atomic Research Centre strongly linked to Banaras Hindu University and Indira Gandhi CAR, and weakly linked to Annamalai University.

Thus, the main groupings from the auto-correlation institution map are reproduced in the five factor matrix, with some additional information provided on the very weak linkages (especially for Indian Institute of Technology).

While this section and the previous two sections portray institutional linkages from a number of perspectives, they offer little insight as to why the institutions are linked; in particular, what are the technical themes on which the linked institutions collaborate. The next series of results portrays linkages based on commonality of subject matter.

3.2.1.3.4. Institution-phrase cross-correlation map. To display these linkages among institutions more visually, a cross-correlation map was generated (using the TechOasis software) that shows institutional relationships based on the usage of common terminology (Fig. 4).

One immediately observable difference between the institution auto-correlation map and the institution-phrase cross-correlation map is the larger number of displayed linkages and strength of the linkages on the cross-correlation map. The institutional collaboration structure has some significant

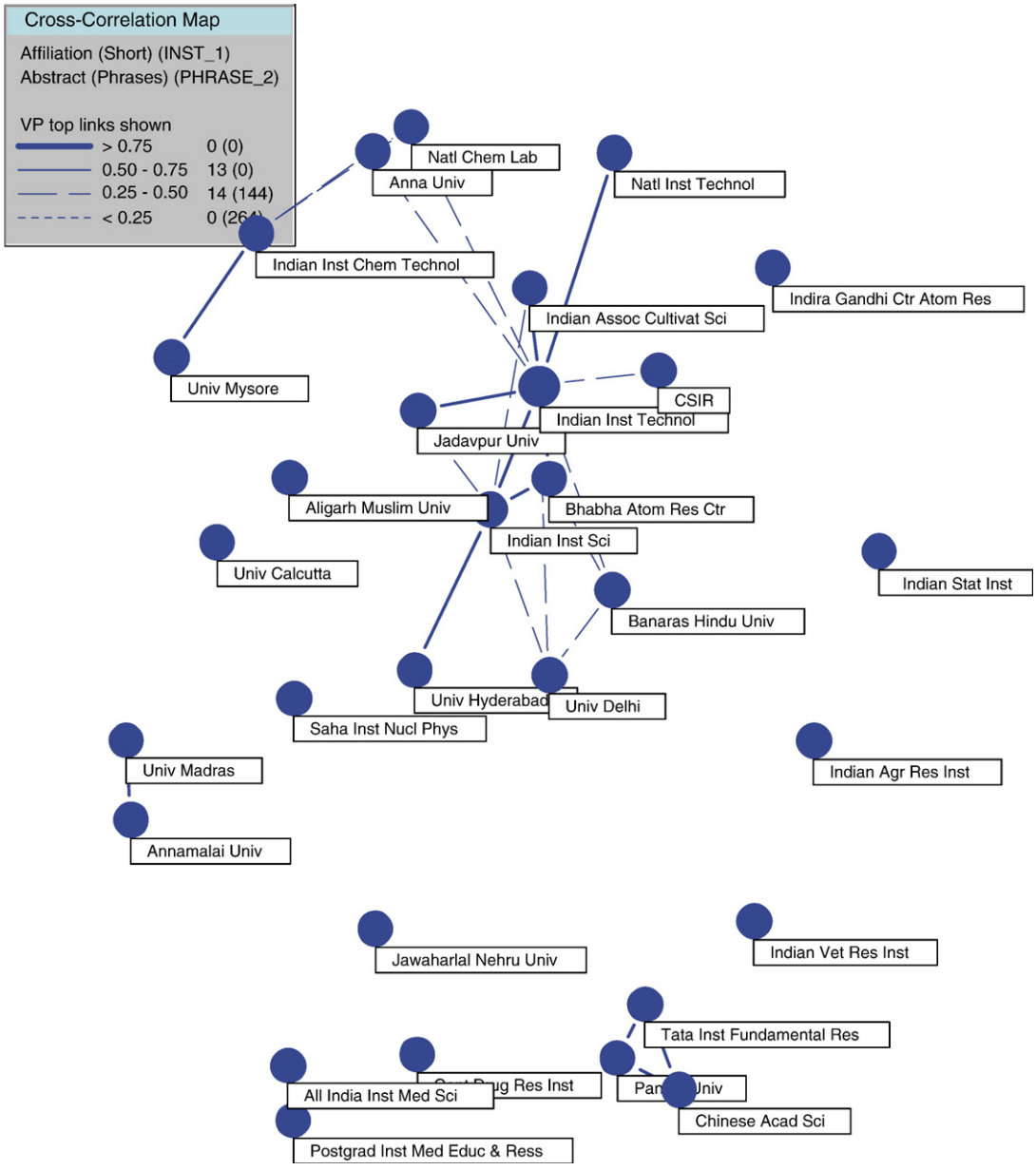

Fig. 4. Institution-Phrase Cross-Correlation Map.

differences from the collaboration structures shown previously. Most importantly, Fig. 4 shows a bi-polar central core of Indian institutional research based on common terminology, with the more basic research centered about the Indian Institute of Science (e.g., proteins, crystals, microfilms) and the more applied research centered about the Indian Institute of Technology (e.g., flow, simulations, macrofilms).

One interpretation of the difference between the structure on Fig. 4 and the previous structures is that the Indian Institute of Science and the Indian Institute of Technology are working in the same general

research areas as a number of other institutions, but they are not collaborating on publications to the same extent. This may be due to overlapping at a generic level of technical description, but distinctness at the much more detailed level of technical description required for collaborative research and publication. Or, it may be due to a tradition of more independent research and publication practices. A more detailed examination of the collaborative practices among the institutions located in the core structure of Fig. 4 might prove fruitful and cost-effective. This approach of comparing institution auto-correlation maps with institution cross-correlation maps may prove to be a powerful approach for identifying institutions that are related by common interests, but are not collaborating accordingly. This auto/cross-correlation map comparison approach need not be limited to institutions. It is equally applicable to authors, countries, and other categories.

From Fig. 4, a few other technically-based groupings can be discerned. There is a medical group at the bottom center (All India IMS, Postgraduate Inst MER, Central Drug Res Inst) that includes a focus on infections, the high energy physics group (common terminology of Belle Detectors, Fermilab Tevatron Collider) at the bottom right identified previously (Tata IFR, Panjab University), a chemistry-oriented group at the upper left (Indian Institute of Chemical Technology, University Mysore, Anna University, National Chemical Lab) emphasizing catalysis and crystal structures, and a medical lab experiment group at the lower left (University of Madras, Annamalai University) that includes an emphasis on animal experiments for liver problems.

On this cross-correlation map, the strong linkages among the Chinese Academy of Sciences, Tata IFR, and Panjab University become more evident. Further investigation into this tripartite relationship shows the involvement of a large number of countries and institutions in a series of high energy experiments conducted at two institutions. The larger effort (in terms of numbers of papers retrieved involving Tata and Panjab) involves the Belle Collaboration, an experiment at the KEK B-factory (Japan) whose goal is to study the origin of CP violation (in particle physics, violation of the combined conservation laws associated with charge conjugation (C) and parity (P) by the weak force, which is responsible for reactions such as the radioactive decay of atomic nuclei). The smaller effort (in terms of numbers of papers retrieved involving Tata and Panjab) involves the high-luminosity Fermilab Tevatron Collider (USA), an experiment whose goal is to search for new particles.

To study the publication dynamics of these three institutions in more detail, all the papers that included an author from Panjab University and Tata IFR in 2004–2006 (time frame expanded to generate better statistics) were retrieved. Many of these papers had tens or hundreds of authors, meaning the institutions (and countries) involved and their researchers were co-authors on many of the retrieved papers. The auto-correlation and cross-correlation maps were difficult to interpret, since the groupings consisted of very large almost concentric circles (reflecting the large overlap in authors and institutions).

For example, Fig. 5 is an institution-phrase cross-correlation map of the Tata-Panjab retrieved papers. There appears to be a large grouping of institutions represented by large highly overlapped circles at the upper right, a smaller grouping of very large highly overlapped circles at the mid-right, and a smaller grouping of mid-sized highly overlapped circles at the lower left. Except for a few instances, identification of specific institutions and their relationships is a challenge.

To clarify and quantify the relationships among these institutions, a factor matrix of institutions from this retrieval (Fig. 6) has been generated. Two factors were selected, based on the main groupings displayed on the auto-correlation map. Each factor represents at least one group of institutions that publish together (and sometimes two groups), and the matrix entries reflect the strength of the publishing

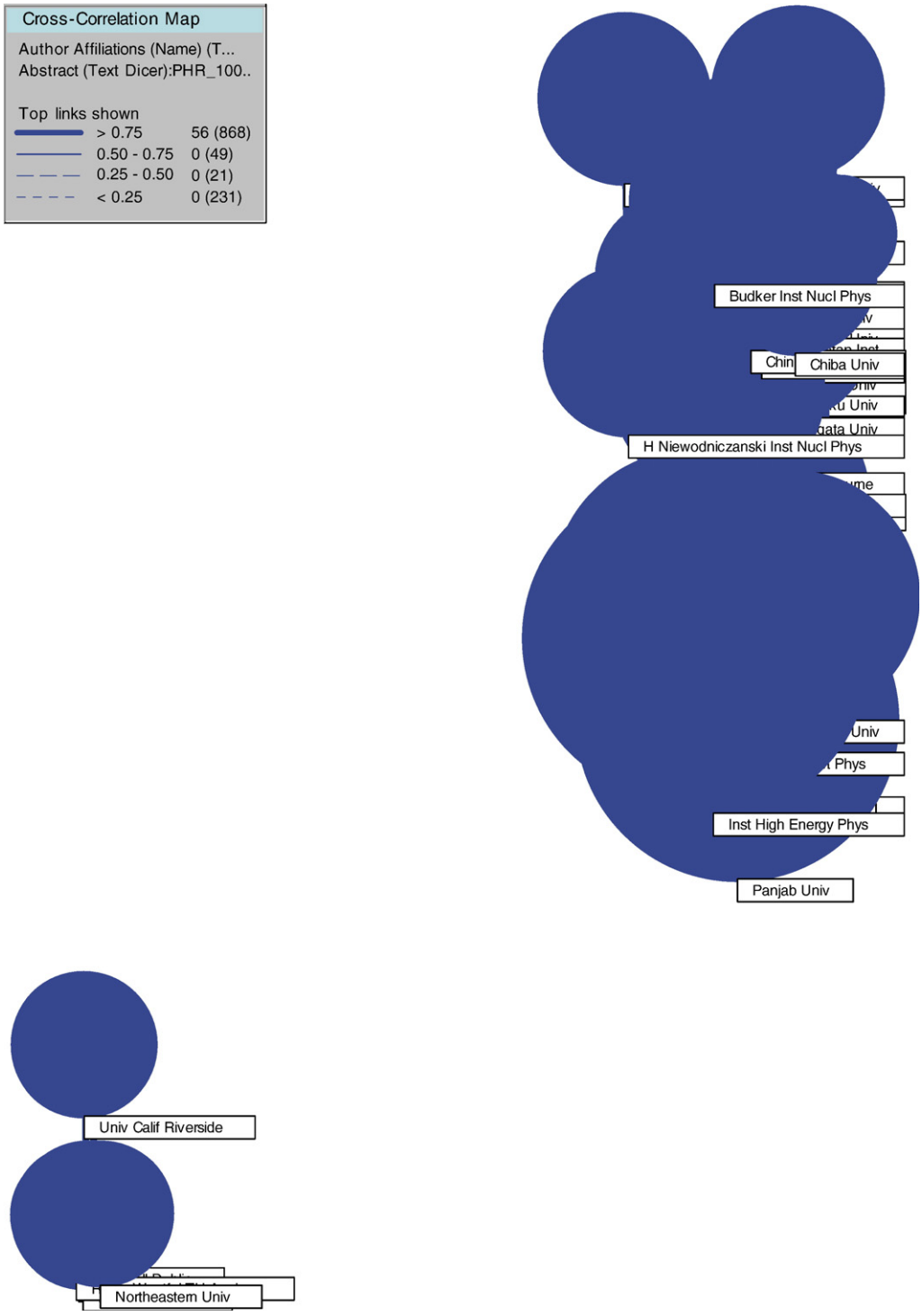

Fig. 5. Institution-Phrase Cross Correlation Map for Taka-Panjab Papers.

relationships. The block nature of some of the publishing relationships becomes evident from the number of institutions having high factor loadings and similar high factor loading values in some of the factor groups.

The shaded cells reflect high factor loadings. Factor 1 shows two main highly correlated groups. The lighter shaded group represents the Belle Detector collaborating institutions, and the darker shaded group represents the Fermilab Tevatron Collider collaborating institutions. Factor 2 shows approximately the same two groups, with the shadings reversed. The Chinese Academy of Sciences is shown as highly correlated with the Belle Detector group. This was validated in a cross-check with the papers retrieved from the SCI, where it was shown that the CAS appeared on two orders of magnitude of more Belle papers than Tevatron papers.

The large group on the upper right of the cross-correlation map (Fig. 5) is the Belle Detector Collaboration (common phrases from the map: Belle Detector; branching fractions; resonance), whose details can be obtained from the lighter shaded group in Factor 1. The smaller group on the lower left of the cross-correlation map is the Fermilab Tevatron Collider collaborators (common phrases: Fermilab Tevatron Collider; production cross sections; quarks), whose details can be obtained from the darker shaded region of Factor 1. The institutions not shaded consist of organizations that contribute to every, or almost every, paper, and therefore participate in both experiments. These include Tata IFR, Panjab University, Princeton University, Institute of High Energy Physics, Institute of Theoretical and Experimental Physics, and Korea University.

There are two main conclusions that can be drawn from this data. First, to understand the richness of the international involvement of the Indian research institutions, some type of network-based analysis will be required to show the complexities. Second, this type of large collaboration effort raises the question as to how the contribution of any one institution or researcher to the total effort would be estimated. However, even though the Indian contribution to the total effort may be small, participation does provide India a 'seat at the table' of an important high energy physics research area, thereby increasing national awareness of potentially significant global technological advances.

3.2.1.4. Collaborative countries. In March 2006, the SCI was accessed to identify the main collaborating countries with India on research articles, in the period 2004–2005. The results are as follows. The format is the name of the country, followed by the number of articles that contained at least one country author and one Indian author. India (46,483), USA (3194), Germany (1441), Japan (1067), England (872), France (711), China (669), South Korea (553), Canada (435), Italy (419), Australia (387), Russia (316), Spain (268).

What is the citation impact of collaboration? Two cases were compared. The first case consisted of all research articles in the SCI published from 1995–1999 having at least one author with an India address. The second case consisted of all Indian research articles in the SCI published from 1995–1999 that excluded India's major collaborators.

The first case (India and collaborators) produced the following results:

- Articles retrieved, 76,717;
- Median citations of total articles retrieved, 2;
- Median citations of top ten cited articles retrieved, 453;
- Median citations of top 5% articles retrieved, 29.

Factor	1	2
Korea Univ	-0.881	0.111
Inst Theoret & Expt Phys	-0.818	0
Sungkyunkwan Univ	-0.662	0.531
Inst High Energy Phys	-0.586	-0.112
Natl Taiwan Univ	-0.5	0.863
Toho Univ	-0.5	0.863
Univ Sydney	-0.5	0.863
Virginia Polytech Inst & State Univ	-0.5	0.863
Tohoku Gakuin Univ	-0.498	0.843
H Niewodniczanski Inst Nucl Phys	-0.497	0.854
Budker Inst Nucl Phys	-0.497	0.853
Univ Maribor	-0.496	0.845
Niigata Univ	-0.496	0.845
Univ Hawaii	-0.495	0.845
Tokyo Metropolitan Univ	-0.495	0.844
Kanagawa Univ	-0.495	0.812
Univ Melbourne	-0.489	0.849
Yonsei Univ	-0.486	0.839
Nara Womens Univ	-0.486	0.838
Nagoya Univ	-0.481	0.866
Tohoku Univ	-0.481	0.866
Univ Cincinnati	-0.478	0.856
Chinese Acad Sci	-0.478	0.796
Osaka Univ	-0.475	0.793
Univ Tsukuba	-0.471	0.798
Tokyo Inst Technol	-0.47	0.836
Tokyo Univ Agr & Technol	-0.469	0.816
Peking Univ	-0.468	0.839
Chiba Univ	-0.468	0.806
Univ Tokyo	-0.457	0.856
Seoul Natl Univ	-0.457	0.796
Univ Ljubljana	-0.448	0.85
Jozef Stefan Inst	-0.448	0.719
Hiroshima Inst Technol	-0.424	0.654
Osaka City Univ	-0.409	0.86
Univ Sci & Technol China	-0.394	0.246
EPFL	-0.335	0.622
Univ Amsterdam	0.454	-0.869
Univ Calif Riverside	0.476	-0.867
Northeastern Univ	0.476	-0.867
Univ Michigan	0.476	-0.867
Rhein Westfal TH Aachen	0.407	-0.811
Univ Lyon 1	0.429	-0.809
Univ Coll Dublin	0.385	-0.771
Tata Inst Fundamental Res	0	0
Panjab Univ	0	0
Princeton Univ	-0.145	0.106
Natl Cent Univ	0.438	0.628
Chonnam Natl Univ	-0.278	0.687
Kyungpook Natl Univ	0.251	0.809

Fig. 6. Factor Matrix of Retrieved Panjab-Tata Institutions.

Table 10
Top Ten Collaborations between India and other Countries for Selected Years

1980		1985		1990		1995		2000		2005	
Country/ Territory	Record Count	Country/ Territory	Record Count	Country/ Territory	Record Count	Country/ Territory	Record Count	Country/ Territory	Record Count	Country/ Territory	Record Count
India	10605	India	10631	India	11560	India	12599	India	16189	India	25209
USA	183	USA	346	USA	416	USA	627	USA	1096	USA	1745
England	54	Fed Rep Ger	101	Germany	120	Germany	203	Germany	437	Germany	795
Canada	45	Canada	94	England	106	England	162	Japan	301	Japan	587
Fed Rep Ger	41	England	92	Canada	88	France	147	England	288	England	503
Japan	23	Japan	52	Japan	65	Canada	125	France	229	France	419
Italy	20	France	35	France	54	Japan	105	Canada	151	South Korea	322
Australia	15	Italy	26	Italy	41	Italy	102	Italy	138	Peoples R China	310
France	15	Australia	25	Switzerland	34	Russia	65	Peoples R China	121	Canada	248
Switzerland	13	Switzerland	18	Australia	29	Switzerland	53	Australia	94	Italy	222
Sweden	12	Sweden	17	Netherlands	28	Australia	48	Russia	88	Australia	215

The second case (India only) produced the following results:

- Articles retrieved, 66,896;
- Median citations of total articles retrieved, 2;
- Median citations of top ten cited articles retrieved, 212;
- Median citations of top 5% articles retrieved, 24.

Thus, approximately 15% of the research articles having at least one author with an India address were the result of India's collaboration with other countries. The impact of collaboration was negligible on median citations of the total retrieval. The impact of collaboration was substantial on the top ten cited articles, and was noticeable on the top 5% of cited articles.

The top ten collaborative countries with India as a function of time are shown in Table 10. Collaboration with the USA (which is the top collaborator) has increased from about 2% of the total articles in 1980 to about 7% in 2005. Collaborations with other nations have also grown steadily, with Germany, Japan, England, and France being other top collaborators.

3.2.1.4.1. Citation bibliometrics. The second group of metrics presented is counts of citations for papers published by different entities.

3.2.1.4.2. Citation trends over time. The numbers of citations of papers with at least one Indian author are presented in Table 11. For almost two decades, growth appears frozen. Only in recent years has growth increased noticeably. However, the number of papers with more than 100 citations appears to be outpacing total publication growth. This is also reflected in the steadily increasing median of the top twenty cited articles. The median of the top 1% has been stable (the 2000 results may not have had sufficient time to garner essentially the remainder of their eventual citations), as has been the overall median.

3.2.1.4.3. Characteristics of most cited vs least cited papers. The papers with the most citations (greater than 150) were compared with those with the least citations (zero) for two separate time periods (1979–1987, 1998–2003), in order to track the changes in characteristics of these papers. The papers with zero citations were obtained by random sampling of all articles published with zero citations in the time frame of the most cited articles. The results are presented in Tables 12–15.

Table 11
Citations of Papers with Indian Authors for Selected Years

Year	Number of papers	Papers with more than 100 cites	number of cites		
			med of top 20	med of top 1%	Overall median
1980	10606	16	117	59	2
1985	10632	17	130	66	2
1990	11563	19	145	68	3
1995	12603	29	151	69	3
2000	16197	22	128	55	2
2005	25227	0	29	10	0

Note: the data for 2005 reflects the short period of time elapsed since the 2005 papers were available to be cited.

Table 12A displays the highest frequency journals with Indian-authored articles having greater than 150 citations, and Table 12B displays the journals containing the zero cited articles. Tables 13, 14 and 15 show the authors, institutions, and collaborative countries, respectively, that are associated with Indian-authored papers that have greater than 150 citations and zero citations.

Table 12

A. Journals Containing Most Cited Papers (>150 citations) Published in 1979–1987

Freq(>1)	Most Cited — 1979–1987 Journals
2	American Journal of Medicine
2	Biotechnology and Bioengineering
2	Contributions to Mineralogy and Petrology
2	Journal of Molecular Biology
2	Journal of Physical Chemistry
2	Journal of Solid State Chemistry
2	Journal of the American Statistical Association
2	New England Journal of Medicine
2	Nucleic Acids Research
2	Transactions of the Institution of Chemical Engineers

B. Journals Containing Least Cited Papers (zero citations) Published in 1979–1987

Freq(>1)	Least Cited — 1979–1987 Journals
3	Indian Journal of Agricultural Sciences
3	Indian Journal of Technology
2	Indian Veterinary Journal

C. Journals Containing Most Cited Papers Published in 1998–2003

Freq(>1)	Most Cited — 1998–2003 Journals
8	Journal of High Energy Physics
8	Nature
7	Physical Review Letters
5	Science
2	Circulation
2	European Physical Journal C
2	Lancet
2	Physical Review D

D. Journals Containing Least Cited Papers Published in 1998–2003

Freq (>1)	Least Cited — 1998–2003 Journals
4	Indian Journal of Agricultural Sciences
3	Indian Veterinary Journal
2	Annals of Arid Zone
2	Asian Journal of Chemistry
2	Bulletin of Materials Science
2	Iete Journal of Research
2	Journal of Applied Animal Research

Table 13

A. Authors of Most Cited Papers Published in 1979–1987

Most Cited — 1979–1987	
Freq(>1)	Authors
3	Ghose, TK
3	Kennard, O
3	Mehta, CR
3	Patel, NR
3	Shakked, Z
3	Viswamitra, MA
2	Chopra, KL
2	Cruse, WBT
2	Ganguly, P
2	Joshi, CP
2	Joshi, JB
2	Khuroo, MS
2	Mukherjee, D
2	Rabinovich, D
2	Salisbury, SA
2	Tyagi, RD

B. Authors of Least Cited Papers (zero citations) Published in 1979–1987

Least Cited — 1979–1987	
Freq(>1)	Authors
(Many authors with frequency of unity)	

C. Authors of Most Cited Papers Published in 1998–2003

Most Cited — 1998–2003	
Freq(>1)	Authors
9	Sen, A
6	Kang, JH
6	Kim, HJ
6	Matsumoto, T
6	Watanabe, Y
5	Eidelman, S
5	Nagasaka, Y
5	Tanaka, Y
4	Kumar, S
4	Lebedev, A
4	Takahashi, T

D. Authors of Least Cited Papers Published in 1998–2003

Least Cited — 1998–2003	
Freq(>1)	Authors
2	Kumar, S
(Many authors with frequency of unity)	

From 1979–1987, the journals containing the most cited papers (Table 12A) were well-known international journals, especially in biology, chemistry, and medicine. Conversely, the journals containing the least cited papers (Table 12B) were predominantly domestic Indian journals, along with other domestic European journals. While the journals containing the least cited papers included journals with a basic research focus, most appear quite applied. There was substantial representation from agricultural and veterinary sciences.

The journals containing the most cited papers in the recent time frame (Table 12C) are well-known physics, multidisciplianry and medical journals. There is at least one anomaly. All the highly cited articles published in the Journal of High Energy Physics are due to one researcher (Dr. Sen), writing on tachyons. In the same time frame, the journals containing the least cited papers (Table 12D) tend to be domestic journals, and include agricultural and veterinary journals as well.

From 1979–1987, the authors associated with the most cited papers (Table 13A) tended to have mainly Indian names, along with a few Anglo names. The authors associated with the least cited papers (Table 13B) all had Indian names. In the recent time frame, authors associated with the most cited papers (Table 13C) included many non-Indian names (especially Japanese names). These results were skewed by Indian participation in two high energy physics research projects, one of which had heavy Japanese participation. Hence, the presence of Japanese names with similar large frequencies in Table 13C. These large multi-national research projects tend to have many (sometimes hundreds) participants, and tend to list all (or almost all) of the participants on each paper. The least cited papers in this time frame (Table 13D) are associated with all Indian names.

In the 1979–1987 time frame, institutions associated with the most cited papers (Table 14A) are dominated by the two most prolific Indian institutions. A few non-Indian institutions are also listed. Institutions associated with the least cited papers (Table 14B) are all Indian.

In more recent times, institutions associated with the most cited papers (Table 14C) are mainly non-Indian. This is further evidence of the high energy physics anomaly discussed previously, since the institutions listed (with similar high frequencies) tend to be high energy physics-oriented institutions. The least cited papers (Table 14D) are again from all Indian institutions.

In the 1979–1987 time frame, the USA was the dominant collaborator on highly cited papers (Table 15A), appearing on almost 28% of these highly cited papers. The other significant contributor was England. India was the only country represented on the least cited papers (Table 15B). In recent times, the USA continues its role as most significant collaborator (Table 15C), appearing on about 2/3 of the highly cited papers. Additionally, Germany has become a significant collaborator, appearing on about 1/3 of the highly cited papers. The main country listed on the least cited papers is still almost exclusively India, with a handful of other countries represented (Table 15D). Thus, the USA, which according to Table 10 collaborates on about 7% of India's total papers in recent years, appears an order of magnitude more frequently on the highly cited papers, and proportionately on the poorly cited papers.

4. Taxonomies — document clustering

This section presents the pervasive technical themes of India's research, the relationships among those themes, and the levels of emphasis (number of research articles published) associated with each

Table 14

A. Institutions Producing Most Cited Papers (>150 citations) Published in 1979–1987

Most Cited — 1979–1987	
Freq(>1)	Institutions
10	Indian Inst Technol
9	Indian Inst Sci
3	Dana Farber Canc Inst
3	Indian Inst Management
3	Univ Cambridge
3	Univ Hyderabad
2	All India Inst Med Sci
2	Childrens Hosp
2	Ciba Geigy
2	Imperial Coll
2	Indian Council Med Res
2	Natl Chem Lab
2	Univ Bern
2	Univ Bombay
2	Univ Delhi

B. Institutions Producing Least Cited Papers (zero citations) Published in 1979–1987

Least Cited — 1979–1987	
Freq	Institutions
3	Indian Inst Technol
2	Agra Coll
2	Banaras Hindu Univ
2	Chandra Shekhar Azad Univ Agr & Technol
2	Jawaharlal Nehru Univ
2	Sardar Patel Univ
2	Univ Allahabad
2	Univ Delhi

C. Institutions Producing Most Cited Papers Published in 1998–2003

Most Cited — 1998–2003	
Freq(>1)	Institutions
9	Univ Tokyo
8	Inst High Energy Phys
7	Mehta Res Inst Math+ACY−Math Phys
7	Univ Munster
7	Univ Tsukuba
6	Brookhaven Natl Lab
6	Florida State Univ
6	Korea Univ
6	Kyoto Univ
6	Tata Inst Fundamental Res
6	Tohoku Univ
6	Tokyo Inst Technol

(continued on next page)

Table 14 (*continued*)

C. Institutions Producing Most Cited Papers Published in 1998–2003	
Most Cited — 1979–1987	
Freq(>1)	Institutions
6	Univ Calif Berkeley
6	Univ Calif Riverside
6	Yonsei Univ
5	All India Inst Med Sci
5	Budker Inst Nucl Phys
5	Columbia Univ
5	Iowa State Univ
5	KEK
5	Nagasaki Inst Appl Sci
5	Nagoya Univ
5	Penn State Univ
5	Princeton Univ
5	Univ Arizona
5	Univ Frankfurt
5	Virginia Polytech Inst+ACY−tate Univ
D. Institutions Producing Least Cited Papers Published in 1998–2003	
Least Cited — 1998–2003	
Freq(>1)	Institutions
7	Indian Inst Technol
3	Indian Inst Sci
3	Tata Inst Fundamental Res
2	Cent Elect Engn Res Inst
2	Indian Vet Res Inst

of the themes. The general approach used is to group the retrieved records into categories of similar documents, identify the central themes through phrase analysis of the records in each category, and tabulate the number of research articles associated with each category. Many approaches for grouping these records can be used. The present paper uses document clustering, based on favorable results from previous text mining studies. Document clustering is the grouping of similar documents into thematic categories. For background on document clustering, and the specific partitional approach used, see [4].

Clustering was performed on three years' retrievals (1991, 2002, 2005). Only the 2005 clustering results will be reported here. The detailed clustering results for all three databases are contained in [4].

4.1. Document clustering results

In partitional clustering, the number of clusters desired is input, and all documents in the database are included in those clusters. There were 16 clusters run (in the year 2006) for the 2005 retrieval. The algorithm used to generate the 2005 clusters had the capability to generate metrics for each node in

Table 15

A. Countries Producing Most Cited Papers (>150 citations) Published in 1979–1987

Most Cited — 1979–1987	
Freq	Countries
55	India
15	USA
8	England
3	Switzerland
2	Canada
1	Colombia
1	Cuba
1	Fed Rep Ger
1	France
1	Ger Dem Rep
1	Japan
1	Pakistan
1	Peoples R China
1	Philippines
1	Poland
1	Scotland
1	Sudan

B. Countries Producing Least Cited Papers (zero citations) Published in 1979–1987

Least Cited — 1979–1987	
Freq(>1)	Countries
55	India

C. Countries Producing Most Cited Papers Published in 1998–2003

Most Cited — 1998–2003	
Freq(>1)	Countries
59	India
39	USA
19	Germany
12	England
12	France
11	Japan
11	Russia
11	Sweden
10	Peoples R China
9	Australia
8	Canada
6	Netherlands
6	South Korea
6	Taiwan
5	Hungary
5	Italy
5	Poland
5	Spain

(continued on next page)

Table 15 (*continued*)

D. Countries Producing Least Cited Papers Published in 1998–2003	
Least Cited — 1998–2003	
Freq(>1)	Countries
59	India
4	USA
1	Egypt
1	England
1	Germany
1	Malaysia

the taxonomy, and these node metrics are presented in [4] as well. Finally, the algorithm used to generate the 2005 clusters had the capability to generate titles for each record in the sixteen lowest level clusters. Because of length considerations, the >14,000 titles will not be included in this paper, but sample titles are shown in [4].

Fig. 7 contains the first four levels of the hierarchical taxonomy, with each cell in the matrix representing a technical category.

There are four columns in Fig. 7, each column representing a level of the hierarchical taxonomy. The highest level (1) is the leftmost column, and the lowest level (4) is the rightmost column. The number preceding each category heading is the number of records assigned by the algorithm to the category.

In the following discussion, the categories in Levels 1 and 4 in the table will be described. The contents of the categories in Levels 2 and 3 are self-evident from their headings and from the contents of 1 and 4. Bibliometrics for each category in the taxonomy were obtained, but are not reproduced for space considerations.

Level 1 is divided into two categories: Biomedical/Environment (5513) and Physical Sciences/Mathematics (8795). Biomedical/Environment covers biological and medical research, as well as agricultural and environmental research.

Physical Sciences/Mathematics covers physics, chemistry, and mathematics, with a strong emphasis on the physics and chemistry of surfaces.

Level 4 is divided into thirteen categories. They are described in order of their listing in Fig. 7, starting from the top.

- **Plant Biology (807)**

This category focuses on plants and seeds, especially the extraction of oils from seeds, and has a food technology emphasis. It appears quite applied, and the main institutions are agriculture-food focused. Other Asian countries play a role equal to that of the USA, although the relatively small amount of Chinese collaboration is somewhat surprising.

- **Animal Experiments (651)**

This category focuses on laboratory experiments for addressing diseases, especially for testing the impacts of drugs. The two main institutions, University of Madras and Annamalai University, were identified on Fig. 2B as having common interests in liver problems especially, and this category confirms that previous finding.

• **Cell Biology/Genetics (1168)**

This category focuses on cell biology and genetics, especially proteins and gene expression. It is one of the more fundamental research categories, as evidenced by the journals and terminology. As expected, the USA is by far the major collaborator in this fundamental research area.

• **Human Patient Diseases (1218)**

The focus here is clinical patient treatment, with emphasis on treatment of infections, especially HIV. Again, the USA is the major partner, with the Western democracies playing strong roles.

• **Soil/Crop Experiments (952)**

The focus is study of soils and plant genetics to improve crop yields. It is more fundamental than the related Plant Biology category, as evidenced by the major journals, keywords, and institutions. The USA is a more dominant collaborator than in the Plant Biology category.

• **Geological Research/Material Mechanics (717)**

This category has two dis-similar thrusts: geological and associated environmental research, and the mechanics of materials. The common links that resulted in these thrusts appearing in the same category are stresses in solid materials and mechanical properties of materials. The next level of dis-aggregation would probably result in separation of these thrusts into different categories. The geological thrust focuses on sediments, and the materials thrust focuses on welding.

• **Algorithms/Network Modeling (1372)**

Focuses on algorithms and modeling of networks, especially communications. While the USA is the dominant collaborator, China plays a noticeable collaborative role in this technology-oriented category.

• **Continuum Analysis (1255)**

Focuses on equations modeling continuum fields, especially flow fields and wave equations. Emphasizes mechanics, mainly fluid but some solid as well. Again, the USA is the dominant collaborator.

<table>
<tr><td rowspan="10">(5513) BIOMEDICAL; ENVIRON</td><td rowspan="3">(2626) BIOLOGICAL RESEARCH</td><td rowspan="2">(1458) ANIMAL EXPERIMENTS/ PLANT BIOLOGY</td><td>(807) PLANT BIOLOGY</td></tr>
<tr><td>(651) ANIMAL EXPERIMENTS</td></tr>
<tr><td colspan="2">(1168) CELL BIOLOGY/ GENETICS</td></tr>
<tr><td rowspan="3">(2887) CLINICAL MEDICINE; ENVIRON</td><td colspan="2">(1218) HUMAN PATIENT DISEASES</td></tr>
<tr><td rowspan="2">(1669) GEOLOGICAL/ MATíL MECHANICS/ AGRICULTURAL RES</td><td>(952) SOIL/ CROP EXPERIMENTS</td></tr>
<tr><td>(717) GEOLOGICAL RES/ MATERIAL MECHANICS</td></tr>
<tr><td rowspan="6">(8795) PHYSICAL SCIENCES/ MATHEMATIC</td><td rowspan="3">(3691) MATHEMATIC</td><td colspan="2">(1372) ALGORITHMS/ NETWORK MODELING</td></tr>
<tr><td rowspan="2">(2319) MATH ANALYSIS</td><td>(1255) CONTINUUM ANALYSIS</td></tr>
<tr><td>(1064) MOLEC LEVEL CALC</td></tr>
<tr><td rowspan="3">(5104) PHYSICAL SCIENCES</td><td>(2867) SURF PHYS/ CHEM</td><td>(1576) FILM PHYS</td></tr>
<tr><td rowspan="2">(2237) COMPOUND CHEMISTRY</td><td>(1291) FILM CHEM</td></tr>
<tr><td>(939) CHEM BOND/ CRYS STRUCT</td></tr>
<tr><td></td><td></td><td></td><td>(1298) REACT/ CATAL/ SYNTH</td></tr>
</table>

Fig. 7. 2005 Taxonomy — SCI.

- **Molecular Level Calculations (1064)**

Physics-oriented category. Focuses on energy states, and calculations at the atomic and molecular level — strong levels of co-authorship. Basic research is published in more well-known physics journals. Large international high energy physics experiments are involved. USA is dominant, but Germany, Russia, and China play significant roles.

- **Film Physics (1576)**

Category has two main thrusts: small-scale film measurements and film deposition and growth

Small-scale Film Measurements (1166 Records)
Film Deposition and Growth (410)

Focuses on surface physics/films. Main thrusts are small-scale film measurements and film deposition and growth. In both thrusts, USA is eclipsed as a dominant collaborator by an Asian country: Japan for small-scale film measurements, and South Korea for film deposition and growth.

- **Film Chemistry (1291)**

Category has two main thrusts: polymer chemistry/properties and surface wet chemistry

Polymer Chemistry/Properties (479 Records)
Surface Wet Chemistry (812 Records)

Focuses on film chemistry, mainly polymer chemistry/properties and surface wet chemistry. Polymer work appears quite applied. Collaborations in both thrusts quite small are compared with other topical collaborations.

- **Chemical Bonds/Crystal Structures (939)**

Category has two main thrusts: ligand–metal complex synthesis and compound hydrogen bonds

Ligand–metal complex synthesis (460 Records)
Compound hydrogen bonds (479 Records)

Focuses on chemical bonds and crystal structures, emphasizing ligand–metal complex synthesis and compound hydrogen bonds. Ligand–metal synthesis emphasizes domestic journals, and is relatively more applied than hydrogen bond work.

- **Reactions/Catalysis/Synthesis (1298)**

Applied organic chemistry category, emphasizing chemical reactions, catalysis, and synthesis.

5. Research expenditure/output comparison

The purpose of this section is to compare research inputs (i.e., funding) to research outputs (i.e., SCI papers), and determine how well they relate. In making this comparison, it is important that the funding and papers relate to the same budget category and discipline of research.

India's research expenditure by field of science (as per latest available year 2002–03) is given in Table 2 in the Introduction of this Special Issue. The relation between the research expenditures above in Table 2 and the thirteen categories of research output articles above is examined. There were four main categories of expenditures from Table 2 (category/% of budget):

- Natural Sciences (25%)
- Engineering & Technology (55%)
- Medical Sciences (3%)
- Agricultural Sciences (17%).

The thirteen Level 4 research output categories can be classified under the four research expenditure categories in the following very approximate manner (category/# records):

- Natural Sciences (25%)
  - Continuum Analysis (1255) [1/2]
  - Molecular Level Calculations (1064)
  - Film Physics (1576)
  - Film Chemistry (1291)
  - Chemical Bonds/Crystal Structures (939)
  - Reactions/Catalysis/Synthesis (1298)
  - Geological Research/Material Mechanics (717) [1/2]
- Engineering & Technology (55%)
  - Geological Research/Material Mechanics (717) [1/2]
  - Algorithms/Network Modeling (1372)
  - Continuum Analysis (1255) [1/2]
- Medical Sciences (3%)
  - Animal Experiments (651)
  - Cell Biology/Genetics (1168)
  - Human Patient Diseases (1218)
- Agricultural Sciences (17%)
  - Plant Biology (807)
  - Soil/Crop Experiments (952).

The relation between the percentage of expenditures assigned to the four funding categories and the percentage of articles assigned to these same categories is as follows (category name/expenditure percent/article percent):

- Natural Sciences (25%/50%)
- Engineering & Technology (55%/17%)
- Medical Sciences (3%/21%)
- Agricultural Sciences (17%/12%).

There are substantial imbalances shown here, and there are many possible reasons for these differences. A few of these possible reasons are as follows.

5.1. Interpretation and assignment

SCI outputs reflect mainly basic research, some applied research, and a small amount of technology development. These research category definitions reflect the research definition perspectives of those responsible for selecting journals to be accessed by the SCI.

The assignment of the SCI outputs to technical categories is governed by two factors: how the clustering algorithm groups records into different thematic areas, and how the thrusts of these thematic areas are interpreted by the authors.

India has mainly adopted the framework articulated by UNESCO for capturing R&D data through primary survey of S&T units in the country. R&D Statistics [6] is the outcome of that process. We have used the funding statistics as given in this report. We believe it is the most authentic data available. However, using these funding statistics does not allow us to disaggregate further from the four categories. Due to this limitation, we were not able to analyse the correspondence with themes in more detail.

5.2. Fraction of output examined

Funds reflected in budget categories may result in many different types of outputs.

Potential research outputs include: papers in journals accessed by SCI; papers in journals not accessed by SCI; papers at conferences or in conference proceedings; patents; presentations; etc. The fraction of total outputs represented only by papers in SCI journals is unknown.

Biasing of outputs through SCI selection is unknown. For example, some technical budget categories could have a high component of very applied research or development funds. Most resulting publications would not be appropriate for SCI journals but may be published in more applied journals.

In practice, Indian journals not accessed by SCI should also be examined. If these domestic journals are, on average, more applied than SCI journals, and if much of the budget category is devoted to very applied work, combining the domestic non-SCI journals with the SCI analysis might provide a more accurate representation of India's research output and its relation to the funding categories.

Isolating which of the above reasons are most responsible for the funding allocation imbalances is a very complex process, and was beyond the scope of this study.

The primary objective of this study was to examine the structure of India's research at higher levels. Accordingly, the thirteen categories for the 2005 database are at a relatively coarse level of resolution. In particular, somewhat more accurate results relating research outputs to research expenditures (above) would be possible with much more well-defined categories, especially for better alignment of specific technical disciplines. An assessment oriented towards more specific technology analyses would require narrower more well-defined clusters, translated into using a larger number of clusters. The present technique is fully translatable into analyzing hundreds or thousands of clusters.

6. Comparison of India's and USA's investment allocations

To place India's research activity in context, the relative investment allocations of India and the USA, with a resolution at the critical sub-technology level, was undertaken. The details of this approach are described in the Introduction of this Special Issue. The results are shown on Tables 16 and 17.

In Table 16, India's research emphasis sub-areas relative to the USA are in traditional agricultural products (e.g., rice husk, groundnut, linseed oil, chickpea, wheat straw), phenomena in the visible spectrum (e.g., optical band gap, visible region, UV–visible, spectrophotometer*, photocatalyst, photodegradation), and chemical topics (e.g., ammonium sulphate, hydrazine, benzene ring, corrosion resistance). This is a combination of traditional agricultural research, applied chemical research, and more physics-oriented research concentrated visibly.

Conversely, as shown in Table 17, the USA's research emphasis relative to that of India focuses on medical, psychological, and sociological applications.

7. Summary and conclusions

India's research article production reached a plateau during the period 1980–1995, and it has since started a rapid increase.

The main technical focus at present is on the three physical sciences in areas of chemistry, physics, and materials, supported by a strong foundation of applied mathematics.

Half the journals that contain most of the Indian papers are domestic Indian journals, and they have low Impact Factors.

Table 16
Indian Strengths

DMWord	USA 2006 Abstracts	India 2006 Abstracts	ABS Ratio India/USA	Normalized Ratio India/USA
Ammonium Sulphate	1	26	26	281.0945
Rice Husk	2	18	9	97.30193
Marker Enzyme*	4	34	8.5	91.89627
Groundnut	12	50	4.166667	45.04719
"Beer–Lambert*" or Beer's Law	16	62	3.875	41.89389
Linseed Oil	3	11	3.666667	39.64153
Optical Band Gap*	16	51	3.1875	34.4611
Chickpea	13	34	2.615385	28.27578
Microwave Irradiation	53	126	2.377358	25.7024
Medicinal Plant*	51	97	1.901961	20.56272
Mulberry	7	12	1.714286	18.5337
Coconut*	25	37	1.48	16.00076
Non-Linear Optical	17	24	1.411765	15.26305
Visible Region	32	44	1.375	14.86557
Wheat Straw	22	30	1.363636	14.74272
Hydrazine	55	70	1.272727	13.75987
XRD	431	537	1.24594	13.47026
Silkworm	16	19	1.1875	12.83845
UV–Visible	91	96	1.054945	11.40535
Thermogravimetr*	193	202	1.046632	11.31548
Spectrophotometr*	228	238	1.04386	11.28551
Benzene Ring*	45	42	0.933333	10.09057
Corrosion Resistance	56	47	0.839286	9.073791
Photocatalyst	27	18	0.666667	7.207551
Photodegradation	51	28	0.54902	5.93563

Table 17
USA Strengths

DMWord	USA 2006 Abstracts	India 2006 Abstracts	ABS Ratio USA/India	Normalized Ratio USA/India
Terrorist*	199	1	199	18.4067
Insurance	858	6	143	13.22693
Abuse or Abusive	1990	14	142.1429	13.14765
Attitudes	1775	13	136.5385	12.62926
Physicians	1891	14	135.0714	12.49357
Mental Health	1397	12	116.4167	10.76808
Prostate Cancer	1757	22	79.86364	7.387067
Adolescen*	3422	43	79.5814	7.360961
Anxiety	1588	20	79.4	7.344182
Immunohistochemical	990	13	76.15385	7.043926
Colorectal	1277	20	63.85	5.90587
Disabilit* or Disabled	1759	29	60.65517	5.610361
Heart Failure or Cardiomyopathy	1928	34	56.70588	5.245067
Depressi*	3955	76	52.03947	4.813443
Lung Cancer	1377	27	51	4.717296
Ethnic*	2130	50	42.6	3.94033
Immunofluorescen*	620	15	41.33333	3.823168
Anestheti*	908	23	39.47826	3.651581

Collaboration with external researchers has the effect of dramatically increasing the absolute number of papers and presence of papers with Indian authors in the higher Impact Factor journals, as well as the numbers of papers with high citations.

India is increasing its growth of articles in highly cited journals greater than its overall increase in growth of research articles overall.

Journals most cited by Indian authors are all international, and have Impact Factors of an order of magnitude larger than the Impact Factors of the journals that publish most of the Indian papers.

The network of Indian co-publishing institutions is weakly linked, but the network of institutions with common thematic interests has some very strong links. In particular, the Indian Institute of Technology and the Indian Institute of Science (India's two leading research publishing institutions) have no strong co-publishing links, but they are at the centers of a bi-polar core of the network of institutions with common thematic interests.

In comparing the most cited Indian papers with the least cited, the following characteristics were identified:

- The most cited were published in international journals, while the least cited were published in domestic Indian journals. The most cited emphasized chemistry, physics, and medicine, while the least cited had substantial representation from agricultural and veterinary sciences.
- The names of the authors of the least cited were all Indian. The names of the authors of the most cited (published a few decades ago) were mainly Indian, but due to an anomaly, the names of the top authors of highly cited papers published relatively recently are non-Indian.

- Institutions associated with the least cited papers are mainly Indian, and this has not changed with time. For papers published decades ago, institutions associated with the most cited are mainly Indian, and are dominated by the Indian Institution of Technology and Indian Institute of Science. For relatively recent papers, institutions associated with the most cited are non-Indian, due to the same anomaly referenced above.
- The USA has been the leading collaborator on the most highly cited papers for decades, increasing its participation from 28% in 1979–1987 to 65% in 1998–2003. India has remained essentially the only country associated with the least cited papers, with a handful of countries listed with very low frequency in recently published papers.

In comparing India's research investment allocation relative to that of the USA, India's strong relative emphasis was on traditional agricultural products, phenomena in the visible spectrum, and selected chemistry topics. The USA's relative research emphasis areas focused on medical, psychological, and sociological.

References

[1] 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.
[2] World Investment Report 2005, Transnational Corporations and the Internationalization of R&D, UNCTAD (United Nations), 2005.
[3] R.N. Kostoff, J.A. Stump, D. Johnson, J.S. Murday, C.G.Y. Lau, W.M. Tolles, The structure and infrastructure of the global nanotechnology literature, J. Nanopart. Res. 8 (3–4) (2006) 301–321.
[4] 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/.
[5] EC. Compendex®, Engineering Village, Elsevier Engineering Information, Inc., Hoboken, NJ 07030, 2006.
[6] Research and Development Statistics: 2004–05 (2006), Government of India, Ministry of Science and Technology.

Ronald N. Kostoff received his Ph.D. degree in Aerospace and Mechanical Sciences from Princeton University in 1967. He has worked for Bell Laboratories, the Department of Energy, and the 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.

Dustin Johnson received his B. S. degree in Electrical and Computer Engineering from Cornell University. He currently works for Northrop Grumman's Intelligence Group (NGIT/TASC) as a Communications Engineer and Technical Lead for the Wireless Information Operations IR&D initiative. Mr. Johnson has served as technical consultant and research coordinator in the recent NYC Emergency Response System proposal.

Christine A. Bowles has an M.S. degree in Materials Science and Engineering from Johns Hopkins University. Ms. Bowles has more than 15 years of experience in materials engineering with emphasis on corrosion and electrochemistry. She worked as a Materials Engineer at the Naval Surface Warfare Center (White Oak/Carderock) for almost 10 years, spent close to 4 years in the chemical processing industry, and is currently a Materials Engineer at DDL Omni Engineering.

Sujit Bhattacharya received his Ph.D. degree from the Indian Institute of Technology in Informatics, and is presently a scientist in the National Institute of Science Technology 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.

Alan S. Icenhour received his M.S. and Ph.D. degrees in Nuclear Engineering from The University of Tennessee. He has 20 years experience with the nuclear fuel cycle, ranging from reactor operations to radiochemical research. He is presently a Technical Advisor on non-Proliferation at the National Nuclear Security Administration (NNSA), on rotational assignment from the Oak Ridge National Laboratory (ORNL), where he is a senior R&D staff member.

Kimberley Nikodym received her B.S. degree in Oceanography from the United States Naval Academy, and a dual M.S. in Oceanography and Meteorology from the Naval Post Graduate School. After serving in a number of command positions in the Navy, she left active duty in 1997. She has worked as a forecaster for the National Weather Service, and presently serves in the naval reserves supporting the U.S. Pacific Command.

Ryan B. Barth received his B.S. degree in Mechanical Engineering from Iowa State University. He currently works for DDL OMNI Engineering as a Mechanical Engineer. Mr. Barth has been performing Textual Data Mining on the transport and distribution logistics literature.

Simha Dodbele received his Masters and Doctorate degrees from the University of Maryland for his work on Vortex flows on delta wings. He worked at NASA Langley Research Center as a contractor, and at NAVAIR on F-18 high lift configurations. He has written over 40 technical publications in aerodynamics, and is an Adjunct Professor at George Washington University and Northern Virginia Community College.