A Bibliometric Overview of Current Publications by Overseas Indians Using Searches Based on Some Common Indian Names

Aparna Basu

aparnabasu.dr@gmail.com
B 303 Nayantara, Sector 7 Plot 8B, Dwarka, New Delhi 110075 (India)

Abstract

Highly educated Indians in science, medicine and engineering have migrated to other countries over several years. Many, currently engaged in research in their adopted countries, could constitute a potential knowledge resource for India that has not been adequately mapped. The country of origin of diaspora authors can be correctly identified through characteristic ethnic names, unique to the country of origin. The large number of ethnic names poses a challenge for complete retrieval, and omissions are inevitable. Using 50 commonly occurring Indian names, we have retrieved a large sample of 23723 Indian diaspora documents indexed in SCI-E for 2007. Names not unique to India were dropped. These 50 names successfully retrieved 53.6% of the SCI-E output from India in 2007. Analyzing by document type, research area, and country/institution of location/collaboration we get a bibliometric overview of current publications by overseas Indians, and compare with broad characteristics of Indian publications. We infer that current publication output of overseas Indians (of which 65% is from the US) could be of the same order as that from India, but with significant differences in the emphasis on research areas. Many authors are either located in, or collaborating with prestigious foreign institutions, primarily in the US.

Introduction

Diaspora: A scattered people, living outside their country of origin

Emigration from India has a long history, beginning with indentured labour in the 19th century to migration of highly educated scientists and engineers, especially to the US from mid 20th century. Even when the initial migrants were uneducated, the second and third generation often achieved distinction in their fields. Earlier, migration was viewed as 'brain drain' from the national perspective [Khadria, 1999, Mahanti, et al, 1998], but later efforts were made by the government to establish contact with émigrés, and initiate programmes to promote exchange with Indian scientists, e.g., TOKTEN (1980-2001) Rao (2006). At present, there is an urgent need to understand the nature of expertise that exists in the expatriate community. A preliminary overview can be obtained through a bibliometric analysis of the published output of overseas Indian scientists. (In this paper, the terms *'Indian diaspora', 'overseas Indians', 'migrants'* are used interchangeably.*)*

Not too many bibliometric studies have been done on ethnic and diaspora groups, a few examples being Lewison (1999) on Serbian and Croatian scientists, Bassecoulard et al. (2003) on Malagasy scientists, Webster (2004) on ethnic minorities in UK science, Basu and Lewison (2006) on genome scientists of Indian origin in US, and Jin, et al (2007) on international collaboration between China and ethnic Chinese scientists overseas. Challenges posed by bibliometric studies of the diaspora relate to correctly identifying authors by country of origin, usually by ethnic names. In this paper, we have used just 50 common Indian names to capture a sizable proportion of the Indian diaspora publications, and analyzed them to understand their characteristics and compare with India.

Data and Methods

An initial list of 50 common Indian names/ surnames or family names was first generated by scanning a large sample of Indian names. The list was used to search for publications by Indians overseas from the Science Citation Index-Expanded (SCI-E) for the year 2007. The

Proceedings of ISSI 2009, pages 697–701.
Edited by B. Larsen and J. Leta.

final list of 45 names used (after removing 5 non-unique ones from the initial list of 50 names, namely, Khan, Shah, Roy, Ray and Thomas) is L1,

(Agarwal Aggarwal Agrawal) Arora Banerjee Basu Bhattacharya Biswas Chakraborty Chandra Chatterjee Das Garg Ghosh Gupta Jain Joshi Kaur Kulkarni Kumar Mandal (Mishra Misra) Mittal Mohan Mukherjee Nair Pal Pandey Patel Prakash Prasad Rai Rao Reddy Sarkar Saxena Sharma Shukla Singh Sinha Srivastava Tripathi Verma Yadav

The union of the sets retrieved from the name searches forms the basic data, which are then analyzed to check for the distribution of publications in the categories 'Document type', 'Subject area', and 'Institutions' using the in-built 'Analyze' module in Web of Science. Comparisons are made between sets of publications retrieved by author searches both in India and overseas. All types of publications, e.g., Articles, Letters, Reviews, etc. have been considered. It is expected that the common names used here will yield a high percentage of publications because the names are likely to occur frequently among authors.

Unfortunately, the searched list of names L1 does not give a good regional distribution of names from India. There is a preponderance of family names or surnames common in the North, East and Western states. The Southern states often do not use a surname, and the wide variation of given names makes them unlikely to appear in a list of common names. Only a few Southern names, Nair, Menon, Reddy, Rao, used as surnames are in the list. However, many authors from the South do appear in the retrieved samples as co-authors, and they are covered to that extent. Some names are common with neighbouring countries like Pakistan, Bangladesh, Nepal, but number of such papers may not be many.

Results

We were able to retrieve 23723 scientific publications from SCI-E using online searches in the Web of Science (WoS) for the single year 2007, with at least one author from list L1, and all authors located outside India. This sample set is called DIASPORA-L1 or **D-L1**. Out of this list, 1474 records (6.2%) did not have a country name in the address field, and may be regarded as possible errors or false positives.

A corresponding search using L1 for publications from India gave 17,555 publications, called INDIA-L1 or **I-L1**. This sample set covers 53.6%, out of a total contribution of 32,726 publications from India in 2007. From the size of the two samples it would appear that the contribution of overseas Indians is sizeable, and roughly as many papers are likely to be authored by overseas Indians as scientists in India. The retrieval efficiency of L1 is good as we were able to retrieve more than half of all papers from India using just 50 names. The percentage retrieval for the diaspora cannot be exactly predicted, but the sample is sufficiently large to justify the broad analysis we have attempted. The average number of papers retrieved per name searched amounted to 390.1 for the Indian papers and 527.2 for the diaspora papers.

Document Types

There are approximately the same number of articles in the two samples **D-L1** and **I-L1,** but the proportion of Meeting Abstracts in percentage terms is almost five-fold in **D-L1**. While articles constitute 82% of the output **I-L1**, they are 58% of the output for **D-L1.**

Table1. Document types retrieved by name search L1 in SCI-E for 2007

INDIA -L1				DIASPORA-L1		
Document Type	Record Count	% of 17555		Document Type	Record Count	% of 23723
ARTICLE 14498		82.59%		ARTICLE	13661	57.59%
MEETING ABSTRACT	872	4.97%		MEETING ABSTRACT	5962	25.13%
PROCEEDINGS PAPER	813	4.63%		PROCEEDINGS PAPER	1519	6.40%
LETTER 536		3.05%		REVIEW	920	3.88%
REVIEW 443		2.52%		EDITORIAL MATERIAL	645	2.72%
EDITORIAL MATERIAL	305	1.74%		LETTER	618	2.61%
CORRECTION 57		0.32%		CORRECTION	254	1.07%
NEWS ITEM	14	0.08%		NEWS ITEM	85	0.36%
BIOGRAPHICAL-ITEM 12		0.07%		BIOGRAPHICAL-ITEM	33	0.14%
REPRINT 3		0.02%		REPRINT	12	0.05%
BOOK REVIEW	2	0.01%		BOOK REVIEW	11	0.05%
				SOFTWARE REVIEW	3	0.01%

Countries

The main countries from which **D-L1** papers emanate are given in Table 2. If there is collaboration between different countries, there is a contribution to each country from that paper.

Table2. Publications in Web of Science 2007 (SCI-E) by overseas Indians, DIASPORA-L1, by country of location/collaboration, and International co-authored papers for INDIA-L1.

INDIA-L1				**DIASPORA-L1**		
Country/Territory	**Record Count**	**% of 17555**		**Country/Territory**	**Record Count**	**% of 23723**
INDIA 17553		99.99%		USA	15628	65.88%
USA 1106		6.30%		ENGLAND	2960	12.48%
GERMANY 406		2.31%		CANADA	1735	7.31%
ENGLAND 329		1.87%		GERMANY	1140	4.81%
JAPAN 304		1.73%		AUSTRALIA	717	3.02%
FRANCE 268		1.53%		JAPAN	619	2.61%
SOUTH KOREA	258	1.47%		FRANCE	510	2.15%
PEOPLES R CHINA	179	1.02%		PEOPLES R CHINA	505	2.13%
CANADA 171		0.97%		ITALY	484	2.04%
AUSTRALIA 150		0.85%		SCOTLAND	402	1.69%

We note from Table 2, that USA dominates the list of countries where overseas Indian publishing scientists are located, or countries they are in collaboration with, accounting for over 65% of the diaspora publication output. The number of publications by **D-L1** in USA is only slightly less than the number from **I-L1**, and is equivalent to half the total scientific output from India, and more than 14 times greater than India's collaborative papers with USA. For most other countries also, the number of diaspora publications exceeds the number of international collaborative papers with India by some factor ranging from 2-9.

Subject Areas

There are major differences in subjects in which papers are published by the diaspora and by Indian scientists, *see* Table 3A and 3B below. India publishes more in the basic sciences, like Chemistry, Physics and also Materials Science, while Life Sciences and Medical Sciences, like Oncology, Cardiac Diseases or Surgery feature prominently among the top diaspora subjects. Complementarities imply that the diaspora can effectively enlarge the knowledge and skill resources available to the Indian people.

Table 3A Top 10 subject areas in which Indian scientists publish (2007)

Subject Area	Record Count	% of 17555
MATERIALS SCIENCE, MULTIDISCIPLINARY	1256	7.15%
CHEMISTRY, MULTIDISCIPLINARY	945	5.38%
CHEMISTRY, PHYSICAL	926	5.27%
CHEMISTRY, ORGANIC	882	5.02%
PHYSICS, APPLIED	852	4.85%
BIOCHEMISTRY & MOLECULAR BIOLOGY	772	4.40%
PHYSICS, CONDENSED MATTER	653	3.72%
PHARMACOLOGY & PHARMACY	588	3.35%
ENVIRONMENTAL SCIENCES	586	3.34%
PHYSICS, MULTIDISCIPLINARY	491	2.80%

Table 3B Top 10 subject areas in which the Indian Diaspora in Science publishes*

Subject Area	Record Count	% of 23723
BIOCHEMISTRY & MOLECULAR BIOLOGY	1806	7.61%
ONCOLOGY 1337		5.64%
CARDIAC & CARDIOVASCULAR SYSTEMS	1115	4.70%
SURGERY 1108		4.67%
GASTROENTEROLOGY & HEPATOLOGY	1052	4.43%
CELL BIOLOGY	1033	4.35%
HEMATOLOGY 1015		4.28%
MATERIALS SCIENCE, MULTIDISCIPLINARY	833	3.51%
PHARMACOLOGY & PHARMACY	801	3.38%
CLINICAL NEUROLOGY	792	3.34%

Institutions

Institutions where Indian scientists could be located or with whom they collaborate are among the prestigious universities and institutions in USA and Canada. The top 10 institutions are shown in Table 4. No institute from other countries features in the list.

Table 4. Top 10 Institutions by location/ collaboration for overseas Indian scientists

Institution Name	Record Count	% of 23723
UNIV TEXAS	932	3.93%
HARVARD UNIV	531	2.24%
UNIV MICHIGAN	407	1.72%
UNIV CALIF LOS ANGELES	381	1.61%
UNIV ILLINOIS	381	1.61%
UNIV PENN	380	1.60%
OHIO STATE UNIV	367	1.55%
MAYO CLIN	353	1.49%
UNIV PITTSBURGH	346	1.46%
UNIV TORONTO	341	1.44%

Summary and Conclusions

Our novel search method for the diaspora, using the most common ethnic Indian names as search terms, is efficient and eliminates extensive name-based searches, at the same time ensuring adequate retrieval. In our earlier study, Basu & Lewison (2006), 5000 names were used. Here, using just 1% of the names (<50 names) we retrieved 53.6% of all Indian papers in 2007. The average document retrieval was 390.1 and 527.2 papers per name searched, for the Indian and diaspora papers respectively. The name list must be enlarged for more complete retrieval. This is currently ongoing. Errors in precision and recall can arise from either inclusion or exclusion of names not unique to India, e.g., Islamic or Christian names, or those occurring in neighbouring countries. The method can be used for other countries and ethnic groups provided they have distinctive names.

Our analysis shows that the Indian diaspora output is not small - currently more than 23,000 papers yearly- and should be seriously taken into account. This includes It is also evident from our analysis of the subject areas that the knowledge pools represented by the Indian and diasporic scientific communities are different and complementary, and this feature can be used in furthering meaningful interaction and exchange between diaspora scientists and India. The considerably larger fraction of Meeting Abstracts authored indicates their working in forefront areas that are currently the subjects of conferences and workshops. For most countries, the Indian diaspora output far exceeds the internationally co-authored papers with India. Institutions where overseas publishing Indians are frequently located, or with whom they collaborate, are primarily in the USA, and are among the most prestigious.

Acknowledgements: To unknown referees for valuable comments and suggestions.

References

Bassecoulard, E., Ramanana-Rahary, S. & Zitt, M. (2003). The ultra-periphery of science: three contrasting views of the Malagasy contribution in terms of domestic research, the diaspora and special topics. In G. Jiang, R. Rousseau & Y. Wu (Eds.) *Proceedings of the 9th International Conference on Scientometrics and Informetrics* (pp. 10-21). Dalian: Dalian University of Technology Press.

Basu, A. & Lewison, G. (2006). Visualization of a Scientific Community of Indian origin in the US: A case study of Bioinformatics and Genomics, in *Proceedings of the International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting*,10-12 May, LORIA-INIST, Nancy, France

Jin, B., Rousseau, R., Suttmeier, R. P., & Cong Cao (2007). The role of ethnic ties in international collaboration: The Overseas Chinese Phenomenon. Paper presented at the International Conference on Scientometrics and Informatics, Madrid, Spain, *Proceedings of the ISSI 2007*, D.Torres-Salinas & H.F. Moed (eds.) CSIC, Madrid, pp. 427-436

Khadria, Binod (1999). *The Migration of Knowledge Workers: Second-generation Effects of India's Brain Drain* (Sage Publications, 1999).

Lewison, G. & Igic, R. (1999). Yugoslav politics, "ethnic cleansing" and co-authorship in science, *Scientometrics* 44 (2): 183-192

Mahanti, S., Krishna, V.V., Haribabu, E., Jairath, V.K. & Basu, A. (1998). *Scientific Communities and Brain Drain*, Gyan Publishing House, New Delhi

Rao, M.K.D., (2006). Gauging success of a project: a case study of the TOKTEN-India Umbrella Project, *Research Evaluation*, Volume 15(3), pp. 175-186

Webster, B. (2004). Bibliometric analysis of presence and impact of ethnic minority researchers on science in the UK; *Research Evaluation*, vol 13(1), pp 69-76