A more refined method to identify core institutes: A case study of biomass research

M. C. Shukla, S. Saksena and M. R. Riswadkar

Core institutes engaged in research on a given topic are those that contribute the most to published literature on that topic. As such, it appears easy enough to identify the core institutes: all one needs to do is to scan one or more abstracting services that cover the research field in question, note the institutional affiliations of authors, and prepare a tally – when the institutes are arranged in decreasing order of the number of papers contributed by them, we get a ranking order. However, in practice, the procedure turns out to be less straightforward.

IDENTIFICATION of major institutions engaged in biomass research the world over, and in India in particular, was one of the objectives of this study. The purpose of this exercise was to identify the institutions that have been actively contributing to the body of literature in biomass, in the period under investigation. Specifically, the top 50 institutions were to be listed and ranked. An exercise of this kind, it is hoped, will help in identifying institutions that are contributing extensively to the advancement of knowledge. The strategy used to achieve this objective is outlined below.

Conventionally, author-affiliation is used to identify major institutes/organizations. However, using author-affiliation as reported by the abstracting services has two disadvantages: (1) each abstract may not always provide complete affiliation (in the present study, the percentage of entries giving complete affiliation was not the same for each service *[Biomass Abstracts (BA), Chemical Abstracts (CA), Energy Abstracts (EA), Energy Abstracts for Policy Analysis (EAPA), Energy Research Abstracts (ERA), Forestry Abstracts (FA), Fuel and Energy Abstracts (FEA), Forest Products Abstracts (FPA), Indian Energy Abstracts (IEA), and Indian Science Abstracts (ISA)]*, it varied between 80 and 97%) and (2) eliminating the same research paper abstracted by two or more services was not possible (at least in the above approach).

The question, therefore, was whether one could use the author-affiliation reported by abstracting service(s) to identify core organizations in any field of interest for a given time span. An attempt has been made in this study to examine this question and explain the rationale behind its solution. Ideally, such a ranking should be done by obtaining exhaustive lists of published literature from, say, 100 prominent institutions. The selection of these institutions could be based on experience in the subject field. This list could then be narrowed down to the required 50 institutions. This is, however, impracticable and time-consuming. An alternative method is to scan all primary sources of literature published in this period. This, too, has its obvious drawbacks. The next choice is to use the information provided by abstracting services. However, using abstracting services to arrive at such a ranking is not as straightforward a task as it seems. The reasons for this are explained below.

Rationale

An abstracting service provides abstracts of only a portion of all the literature published in a given period. The need for, and the extent of, sampling is governed primarily by the policy of the service, and space considerations. For this reason, it is not advisable to use a single abstracting service to identify the core institutions. By considering many services, the coverage is increased but a complete coverage cannot still be guaranteed. However, using many abstracting services leads to a problem, the crux of which is this: the probability that the title of a primary (paper) document will be mentioned along with the complete affiliation of its author(s) is not necessarily the same for each service. This statement is based on the realization that the appearance of the name of the institution from where a paper is contributed is dependent on two factors, both of which are governed by chance. More importantly, the magnitude of these chances is not the same in all services. These two factors can be framed as questions: (1) What is the chance that a paper (bearing the affiliation

---

This study is based on the thesis accepted by the University of Poona, Pune, for the award of Ph D degree, April 1996.

M. C. Shukla is in the Documentation and Information Centre and S. Saksena is in the Energy Environment Interface Group, Tata Energy Research Institute, New Delhi 110 003, India.
M. R. Riswadkar lives at 389/1, Shanwar Peth, Near Suyog Mangal Karyalaya, Pune 411 030, India.

Table 1. Sample selected for the study

Service	Year of publication					
	1982	1983	1984	1985	1986	Total
Samples of abstracts published						
BA	296	308	328	275	262	1469
FA	333	314	324	328	281	1580
FPA	168	143	152	144	111	718
						3767 (41%)
All the abstracts published						
CA	181	204	190	175	140	890
EA	118	91	163	137	72	581
EAPA	60	80	140	102	148	530
ERA	431	409	277	387	374	1878
FEA	159	110	86	94	99	548
IEA	60	71	139	150	48	468
ISA	65	40	25	290	124	544
						5439 (59%)

Table 2. Reporting pattern of complete affiliation in the field of biomass (all bibliographic forms)

Service	Total records	Complete address available	Percentage
BA	1469	1250	85.1
CA	890	773	86.9
EA	581	550	94.6
EAPA	530	465	87.7
ERA	1878	1703	90.7
FA	1580	1362	86.2
FEA	548	–	–
FPA	718	577	80.3
IEA	468	454	97.0
ISA	544	515	94.7

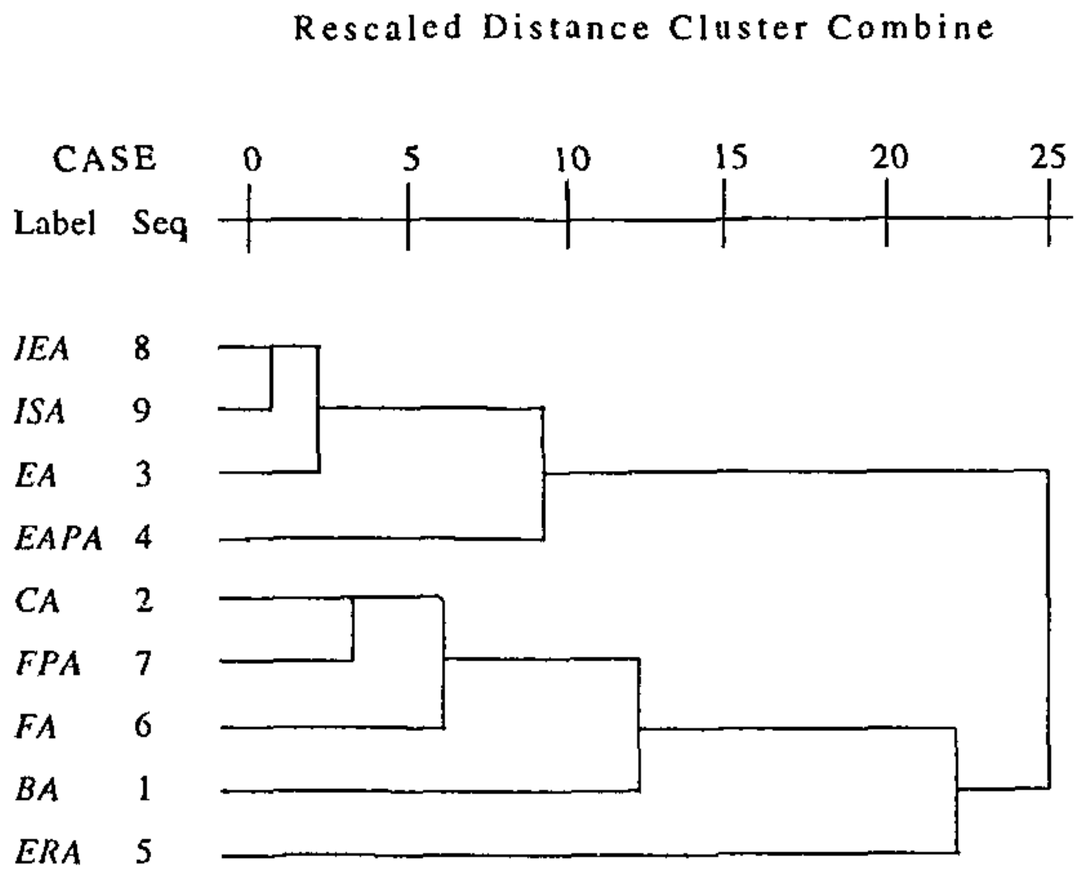

Figure 1. Dendogram using average linkage (between groups).

of the author) will be selected by a service? (2) What is the chance that, having been selected by the service, the complete affiliation will also be printed along with the abstract? If the selection of a paper is not dependent on either the service or the type of paper and if the probability of the complete affiliation of a selected paper being mentioned is the same for each service, this task would be simple: count the number of times each institution is mentioned (after eliminating duplicated records) in the sample of abstracts, disregarding the service that published the abstract. The reality is, however, different.

A simple example can illustrate what has been discussed so far. Suppose an article was published (with full affiliation) by ABC institute. The article was abstracted by two services S and T. Service S is known generally to mention affiliations in 90% of its abstracts, the corresponding figure for service T being 60%. Then it is 1.5 times more likely that one would have noticed the name ABC if one were to consult service S alone. But since abstracts of S and T are pooled, one would notice the name ABC if it is mentioned in S, or in T, or in both. The chance that this will happen is 96%. Now, consider another article published by DEF institute. It is abstracted by services M and N whose affiliation-reporting percentages are 50 and 80 respectively. By similar logic, it can be shown that the probability that DEF will be mentioned in M, or in N, or in both, is 90%. So if all the four services are considered, then the institute ABC is more likely to be mentioned than DEF, the difference in probability being 6%. The situation can get more complicated if one assumes that it is not equally likely that articles from ABC will be abstracted by services S and T. Similarly, one can assume that articles from DEF are not equally likely to be abstracted by services M and N.

There is nothing one can do to avoid the consequences of the fact that a service's coverage can only be selective and that it would be providing affiliations for only a certain fraction of the selected abstracts. But what one can do is to ensure that sets of services are chosen that consist of fairly similar services in these respects. For example, consider that 89%, 91%, and 92% of the abstracts in three services M, N, and P are those of the articles that have appeared in journals. The fraction of abstracts that carry complete affiliation is 75%, 77%, and 76% respectively in these three services. It can then be said that these three services are similar. Therefore, if two institutes – ABC and DEF – publish a journal article each, then it is equally likely that one would encounter their names. It is then immaterial which particular service published the abstract. That is, neither of the institutes will face a handicap. In other words, within a set of services there should be no comparative bias in selecting articles or reporting affiliation.

Methodology

A total of 9206 records were selected as a sample which forms 16.4% of the total output (56,240 abstracts) covering the period from 1982 to 1986 of the ten databases under examination. Stratified random sampling was employed in the case of *BA*, *FA*, and *FPA*. In the remaining seven services (*CA*, *EA*, *EAPA*, *ERA*. *IEA*, and *ISA*) original total (populations) of abstracts were taken for further analysis (Table 1).

Having selected the sample, the question that arises is: How to identify, from the ten services, those that are similar? In the present work, first the fraction of abstracts of journal articles where complete affiliation was provided by each service was ascertained. Overall, about 90% of the abstracts carried complete affiliation. There is no way of knowing how many abstracts in the remaining 10% did not record the affiliation because it was not furnished in the source document itself, and how many failed to do so during the abstracting process owing to human error, negligence, etc. It is assumed, for this exercise, that all primary source documents did carry the complete affiliation. The reporting of complete affiliation varied across the services, from 81 to 97%. A χ^2 test confirmed that reporting fraction was significantly dependent on service. By trial and error, it was possible to isolate two homogeneous sets (confirmed again by the χ^2 test) of services: one reporting affiliations to the extent of 94% and the other, 85%. One service, *EAPA*, still stood out (81%) when only journal articles were considered. It was then decided to consider all bibliographic forms and carry out a similar analysis (Table 2).

The overall reporting fraction in this case turned out to be 89%. The services showed a range from 80 to 97%. Again, it was possible to isolate two homogeneous sets. However, in this instance, *EPA* stood out (80%). It is only a coincidence that in both the exercises, the analysis yielded two sets. This then raised the question: Is use of only one criterion in forming sets of services valid? It was then decided to consider four other criteria: (1) the availability of affiliation in these services, (2) the number of journal titles monitored, (3) the number of conference titles monitored, (4) the number of languages, and (5) bibliographic forms (e.g. B for book, P for patent, J for journal article) considered by each service (Table 3).

It is hypothesized that these five factors primarily determine whether a given document will be selected for abstracting. A statistical technique, cluster analysis[1] was used to group similar services by simultaneously considering all the five criteria. The normal Z-scores of these five variables were used to prepare a 10×5 matrix for cluster analysis. The results showed that the ten services can be sorted into two homogeneous sets (Figure 1). The membership of each of these sets is also decided by cluster analysis. It would have been very fortunate if the results of cluster analysis had shown that all the services were similar and therefore part of a single set. But since

Table 3. Criteria considered for cluster analysis

Service	Availability of affiliation (%)	Journal titles monitored	Conference titles monitored	Languages considered	Bibliographic form(s) covered
BA	85.1	341	240	19	8
CA	86.9	213	69	19	8
EA	94.6	144	45	9	5
EAPA	87.7	78	65	4	8
ERA	90.7	38	218	12	11
FA	86.2	325	98	32	8
FEA	–	118	51	9	7
FPA	80.3	187	34	24	9
IEA	97	51	22	1	4
ISA	94.7	76	4	1	5

Table 4. Top 50 institutes in the field of biomass research

Institute (1)	Number of papers A(2)	C(3)	A′(4)	C′(5)	Rank A(6)	C(7)	A′(8)	C′(9)
Solar Energy Research Institute, Colorado, USA	120	129	144	157	1	2	2	3
Pacific North-West Laboratory, WA, USA	96	97	124	139	2	3	3	4
Institute of Gas Technology, Chicago, USA	71	85	79	107	3	4	4	5
Department of Agriculture, Washington, USA	62	62	221	236	4	6	1	1
Oak Ridge National Laboratory, USA	56	69	58	78	5	5	5	6
University of California, California	37	57	34	61	10	7	12	7
NAEMNDEN, Sweden	45	45	43	43	6	9	6	11
University of Florida, Florida	41	53	41	59	7	8	7	8
StatensEnergiverk, Sweden	40	40	38	44	8	10	8	10
Forest Research Institute, Dehra Dun, India	39	204	38	212	9	1	8	2
South Forest Experimental Station, USDA, USA	37	38	37	38	10	11	9	14
Argonne National Laboratory, USA	33	36	32	45	11	13	13	9
CSIRO, Australia	33	34	31	31	11	15	14	19
North Central Forest Experimental Station, USDA, USA	32	35	36	36	12	14	10	16
Tennessee Valley Authority, TN, USA	32	36	37	38	12	13	9	14
BatteleColumbus Laboratory, USA	32	36	35	45	12	13	11	9
Department of Energy, USA	32	36	31	38	12	13	14	14
Forest Products Laboratory, USA	30	32	27	32	13	16	16	18
Virginia Polytechnic Institute, VA, USA	30	32	27	35	13	16	16	17
Forest Research Institute, New Zealand	29	29	26	26	14	19	17	23
Texas A & M University, Austin, USA	28	37	31	39	15	12	14	13
Lawrence Berkeley Laboratory, Berkeley	27	31	30	32	16	17	15	18
University of Washington, Washington	27	27	25	25	16	21	18	24
CEC, Brussels	26	28	24	27	17	20	19	22
CEC, Luxembourg	25	26	26	28	18	22	17	21
Finish Forest Research Institute, Finland	25	25	22	23	18	23	21	26
Georgia Institute of Technology, Atlanta	23	25	24	24	19	23	19	25
INRA, France	23	29	24	28	19	19	19	21
Pacific Forest Research Centre, Canada	23	25	23	25	19	23	20	24
Swedish University of Agricultural Science, Umea	22	23	35	35	20	24	11	17
Forest & Forest Products Research Institute, Japan	17	17	19	15	25	30	22	32
Idaho University, Idaho	21	21	19	20	21	26	22	28
Auburn University, Auburn	19	19	16	17	23	28	25	31
Cornell University, Cornell	20	26	24	37	22	22	19	15
Indian Institute of Technology, Delhi	7	38	10	39	28	11	27	13
Central Institute of Agricultural Engineering, Bhopal	0	35	0	35	29	14	28	17
University of Arizona, Arizona	14	30	14	29	27	18	26	20
North Carolina University, North Carolina	18	26	16	23	24	22	25	26
Punjab Agricultural University, Ludhiana	0	26	0	25	29	22	28	24
University of Georgia, Georgia	16	20	18	18	26	27	23	30
Central Arid Zone Research Institute, Jodhpur, India	0	23	0	23	29	24	28	26
Illinois University, Illinois	16	23	17	21	26	24	24	27
Kerala Forest Research Institute, Peechi, India	0	22	0	19	29	25	28	29
National Research Council, Canada	19	22	18	24	23	25	23	25
University of Wisconsin, Wisconsin	18	22	16	23	24	25	25	26
Pennsylvania State University, Pennsylvania	19	20	16	25	23	27	25	24
Princeton University, Princeton	17	20	19	23	25	27	22	26
University of Hawaii, Hawaii	18	20	19	23	24	27	22	26
Forest Services, USA	30	38	32	40	13	11	13	12
Texas Tech University, Texas	18	18	18	20	24	29	23	28
Brookhaven National Laboratory, Upton, NY	18	19	18	20	24	28	23	28

that is not the case, the core institutions were identified separately with these two clusters/sets as well as with all of them pooled together.

The members of each group were also decided by cluster analysis. The three groups that were considered are: Group A: *BA, CA, ERA, FA, and FPA*; Group B: *EA, EAPA, IEA*, and *ISA*; Group C, which included all the services (excluding *FEA*). A card was prepared for each paper abstracted containing information on (i) name of the service in an abbreviated form (e.g. *BA, CA*) followed by year of publication and the abstract number, (ii) name of the first author followed by year of publication, (iii) title of document, (iv) source, (v) language of the text, (vi) affiliation of the first author (on

Table 5. Z-values calculated for Wilcoxon matched-pairs signed ranks test, for absolute values

	A	C	A′	C′
A	–	–5.5867 (0.0000)	–1.1293 (0.2588)	–5.4840 (0.0000)
C	–	–	–4.1795 (0.0000)	–2.8453 (0.0044)
A′	–	–	–	–5.5463 (0.0000)
C′	–	–	–	–

Figures in parenthesis indicate two-tailed probability, *p*.

Table 6. Z-values calculated for Wilcoxon matched-pairs signed ranks test, for ranked data

	A	C	A′	C′
A	–	–1.9663 (0.0493)	–0.3135 (0.7539)	–2.2524 (0.0243)
C	–	–	–1.6770 (0.0935)	–2.8117 (0.0049)
A′	–	–	–	–2.7504 (0.0060)
C′	–	–	–	–

Figures in parenthesis indicate two-tailed probability, *p*.

the back of the card), (vii) document type (e.g., B for book; J for journal article; P for patent), and (viii) country of the source document. For each group the cards were arranged first by country and then, within each country, by author. When names of authors were identical, titles of articles were checked. For the same research papers only one card (from any service) was retained and additional cards were kept aside. In the next stage, these cards were arranged by organization under each country. In the last stage, these cards were rearranged in decreasing order of productivity, irrespective of the country. Finally, a ranked list of core institutions (or ranked organizations) was compiled.

Results

The three ranked lists are fairly different. Group B places three Indian institutes first. SERI (Solar Energy Research Institute, Colorado; now known as National Renewable Energy Laboratory) appears in the 19th position. This cannot be the representative situation at the global level. In Group C, Forest Research Institute, Dehra Dun tops the list. Two Indian institutes are ranked at 12th and 19th positions. It is obvious that IEA and ISA have selected mostly Indian abstracts. This led to high ranking of Indian institutions. Group B thus could be used to identify the top 20 Indian institutes. Therefore, it was decided not to consider Group B for final comparative ranking. Table 4 shows comparative ranking for (1) Group A without duplicates, (2) Group C without duplicates, (3) Group A with duplicates (A′), (4) Group C with duplicates (C′).

The results show that comparative ranking does not change significantly if duplicate records are included. However, differences between Group A and Group C are significant. One can either compare the absolute values (columns 2, 3, 4 and 5) or the ranks (columns 6, 7, 8 and 9).

Comparison of absolute values using Wilcoxon matched-pairs signed-ranks test

From Table 5 it is evident that only the groups A and A′ are statistically similar. (Note that for four groups, there are six pairs that can be compared.) Groups are considered to be similar when *Z*-value is low and *p* is high.

Comparison of ranks using Wilcoxon matched-pairs signed-ranks test

By ranking the institutes it is observed from Table 6 that groups yield comparatively more similar results than for the corresponding absolute values (in each cell of Table 6, *Z*-value is lower and *p* value is higher than corresponding cell in Table 5).

It is also observed that the presence of duplicate records does not significantly alter the ranking; i.e. A′ = A and C′ = C. (Of course, duplicates create significant difference in absolute values between A′ & A.) From Table 6 it can be seen that the groups which are most similar are: A & C, A & A′, C & A′.

Conclusions

1. If absolute values of number of publications (the actual frequency distribution) are considered then it matters how one has grouped the services in order to identify major institutions.
2. If ranking is considered alone, then grouping of services is less critical. Also, duplicate records do not affect the ranks (but this is not true of absolute values).
3. The decision to select top 50 institutes out of the four groups, is to be based on experience and intuition, since Table 4 suggests that many of these groups produce similar ranking. Thus, we chose Group A (*BA*, *CA*, *ERA*, *FA*, and *FPA*). If top 20 Indian institutes are to be identified, then Group B (*EA*, *EAPA*, *IEA*, and *ISA*) can be used.

1. McKay, D., Schofield and Whiteley (eds), *Data Analysis and the Social Sciences*, Frances Printers, London, 1983, pp. 226–240.

ACKNOWLEDGEMENTS. We are grateful to Mr Yateendra Joshi, TERI, for his comments and suggestions on an earlier draft of this paper. MCS thanks Dr R. K. Pachauri, Director, TERI, for granting one-year study leave and for the permission to use TERI resources.