

KEYWORD CLUSTERING: A NEW TECHNIQUE TO LOCATE RESEARCH AREA OF A SUBJECT

Kalipada Jana*

* Librarian

Basanti Devi College,
Kolkata, West Bengal,
India

QR Code



Abstract : - Finding out the thrust areas of research is a basic requirement for studying a subject. This study presents a method for identifying research areas in a field by analysing the keywords used in papers. Keywords assigned to 682 research papers published between 2013 and 2017 in a specific subject area, "Black hole," have been collected and analysed. The search phrase "Black hole" was used to obtain the research papers from Web of Science. The assigned keywords occurred with different frequencies. The keywords formed the clusters of words. The cluster names were based on the names of the most common words in each keyword cluster. Based on their size, the clusters have been divided into three categories: small, medium, and large. Since a fairly large number of keywords formed sizable clusters, it is assumed that these clusters represent the potential facets. Three key parameters related to the keyword clusters were identified: the number of keywords in a cluster, frequency of occurrence, and occupancy. Based on these parameters, four indicators were defined: stability index, integrated visibility index, momentary visibility index, and potency index. The fluctuations of these indicators were observed throughout the study period. These indicators have different values for each cluster and facet. Their value ranges are divided into five categories: very high, high, medium, low, and very low. An indicator represents a specific aspect of a facet. These indicators can be used to assess different characteristics of facets.

Keywords: Keyword Cluster; Research Areas; Black Hole; Web of Science, Research Trend, Facet Analysis

1. Introduction:

Identifying research areas is fundamental to studying any subject. As Seetharama (1997)

pointed out, studying the development of a subject involves three main components:

- 1) Landmarks in the subject's development.

- II) Research trends gathered from review documents, indexing and abstracting periodicals, state-of-the-art reports, trend reports, etc. This also includes assessing, in general, (a) the growth of literature (documents) related to the subject and (b) the extent of information dispersion and scattering within the subject.
- III) Trends in education related to the subject at various levels. Research trend analysis mainly focuses on identifying the key facets that represent the subject's main areas of emphasis. This paper proposes a quantitative method to identify facets that constitute the core research areas of the subject "Black HOLES." The method involves selecting keywords from research articles and conducting frequency analysis. This analysis highlights keyword clusters that indicate the subject's active research topics.

2. Literature reviews:

Price (1963, 1975) studied the growth of scientists, scientific journals, and papers over the past two centuries and found that their numbers doubled approximately every 15 years. While literature and information are generally assumed to grow exponentially, growth within individual disciplines may follow a linear pattern. The concept of ageing or obsolescence is closely linked to scientific growth. Some researchers suggest that growth and obsolescence are inversely related, meaning that the faster literature grows in a field, the quicker it becomes obsolete

(Shodhganga, 2015). This relationship can also be explored by analyzing the growth and obsolescence of keywords within a subject. Citation analysis has been used in studies on literature growth and obsolescence as well.

Content analysis is another widely used research method in library and information science to study a subject's current state. Originally applied in mass communication in the 1950s, content analysis is a systematic and rigorous way to analyze research papers generated during a study (Berelson, 1952). Keywords are fundamental elements representing the content of a subject. Content analysis based on keywords assumes that a paper's keywords adequately represent its content. The co-occurrence of two different keywords in the same paper indicates a link between the topics they represent (Cambrosio, 1993).

Several researchers have contributed to keyword-based studies. Coulter, Monarch, and Konda (1998) used keywords selected by professional indexers, valuing the consistency of a fixed taxonomy applied by experts. Looze and Lemarie (1997) conducted co-word studies using expert-proposed keywords. Courtial, Cahlik, and Callon (1994) obtained keywords from online databases to study social interactions, while Law and Whittaker (1992) mapped acidification research through co-word analysis. Noyons and Van Raan (1998) improved science and technology mapping by enhancing graphical user interfaces and adding external map information, helping users focus on

specific areas and actor positions. Their work also included mapping neural networks using classification code co-occurrence. Van Raan and Tijssen (1993) discussed the limitations and potentials of bibliometric mapping via co-word analysis. Van Raan (1997) analyzed major research areas covered in the journal *Scientometrics* using co-word analysis, and Courtial (1994) studied author interaction networks in the same journal.

To analyze temporal variations in a subject's content, the Keyword Associated Content Variation Analysis (KACOVA) model was developed, focusing on detailed keyword abstraction and analysis (Dutta, 2008; Dutta,

Majumder, Sen, 2010). In one study, keywords in the subject "Fermi Liquid" were categorized as keyphrases, modulators, and qualifiers before analysis (Dutta, Majumder, Sen, 2008). A new keyword taxonomy was proposed to analyze research trends and identify potential research areas (Dutta, Majumder, Sen, 2010).

Keywords are commonly used consciously or unconsciously in everyday life. The term "Keyword-in-Context" was introduced by Luhn in 1960 (Luhn, 1960), and early use of keywords was documented in 1975 in the *Journal of Applied Behaviour Analysis* (Hartley and Kostoff, 2003). Keywords represent content, and keyword cluster analysis effectively becomes content analysis of a subject. As a subject becomes more specialized, the words used to describe it are those used by experts in the field—what Luhn calls the "Keyword" or significant word (Black, 1962).

Cluster analysis of keywords is a powerful method for examining how users perceive information spaces, aiming to create flexible and customizable classification schemes. This method relies on statistical analysis of keyword characteristics and is widely applied across disciplines, especially in document-based cluster analysis, which supports automatic classification (Willett, 1988).

A notable limitation of existing subject access tools is their failure to address the behavioral aspects of keywords—that is, how keywords appear and behave in databases, whether full-text or bibliographic. One strength of the model proposed here is its quantitative interpretation of these behavioral aspects, which helps examine the growth and obsolescence of different domains within a subject.

3. Objectives, Methodology and Restriction

The main objective of this paper is to develop a method for identifying research areas within a subject using primary sources that accurately reflect its current state. A total of 143 research articles on "Black Holes" published between 2013 and 2017 were retrieved from the Web of Science. The search term "Black Holes" is assumed to represent a specific subject domain within the broader field of astrophysics. From the titles

and abstracts of these articles, 4,873 keywords were extracted, averaging about seven keywords per article (Table 2).

Year-wise variations in the number of articles and keywords are shown in Figures 1 and 2. The keywords were then divided into two groups: “Clustered keywords” and “Unclassified keywords.” To identify clustered keywords, those sharing common word(s) were grouped together, forming keyword clusters. Clusters were categorized by size based on the number of keywords they contained: “Small clusters” had 3 to 5 keywords, “Medium clusters” had 6 to 10, and “Large clusters” had more than 10 keywords. Clustering keywords essentially groups them into common facets, with each facet defined by a shared word within the cluster. Each cluster is named after its common word. Examples of keyword clusters are provided in Table 1.

Table 1.

Size of the cluster	Name of the cluster and no. of keywords	Keywords of the cluster
Small	Accretion (4)	Accretion; Accretion disc; Accretion efficiency; Advection-dominated accretion
Small	Dark energy (4)	Dark energy; Dark energy theory; Holographic dark energy; New age-graphic dark energy
Medium	Quark (6)	Strange quark matter; Heavy quark physics; Quark production; Quark star; Quark gluon plasma; Quark jet
Medium	Thermodynamics (9)	Thermodynamic geometry; Thermodynamic property; Thermodynamic quantity; Thermodynamics; Generalized law of thermodynamics; First law of thermodynamics; Thermodynamics of FRW universe; Generalized second law of thermodynamics; Thermodynamics of generalized uncertainty principle
Large	Entropy (18)	Entropic force; Entropic acoustic instability; Entropy; Entropy and area quantization; Entropy bound; Entropy conservation; Entropy correction; Bekenstein-Hawking entropy correction; Entropy density; Entropy of horizon; Entropy quantization; Entropy spectrum; Bekenstein-Hawking entropy; Hawking entropy; Canonical entropy; Corrected entropy; Correctional entropy; Entanglement entropy.

Table 2

Year	No of Articles(A)	No. of selected keywords (B)	Total no. of occurrences of all keywords (C)	Average frequency (rounded) of occurrence per keyword (C/B)	No. of selected (rounded) keywords per article (B/A)	No. of occurrence of keywords (rounded) per article (C/A)
2013	117	352	820	2	3	7
2014	147	478	960	2	3	7
2015	129	429	891	2	3	7
2016	143	421	1014	2	3	7
2017	146	465	1188	2	4	8
Total	682	2145	4873	2	3	7

4. Data Presentation and Analysis

The total number of selected keywords over 5 years is 2145 (Table 2), which includes repetitions of the same keywords in different years. The number of distinct keywords over 5 years, i.e., without repetitions, is 371. Hence, on average, each keyword is repeated ($2145/371 = 5.7$) nearly six times over 5 years. On average, three keywords were selected from each article over eighteen years, ($2145/682 = 3.1$) and there are 2.3 ($4873/2145 = 1.27$) distinct keywords per article on average. Out of a total 371 distinct keywords, 112

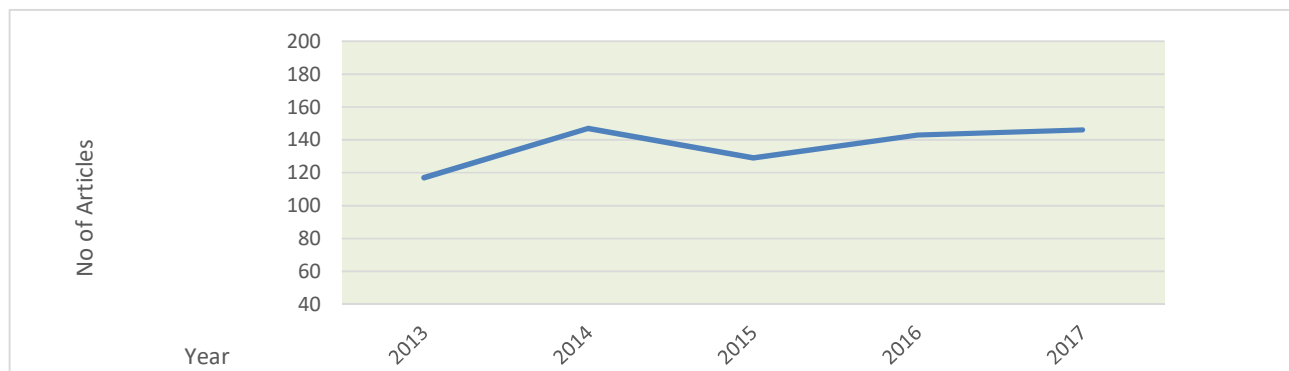


Figure 1. Number of articles over five years.

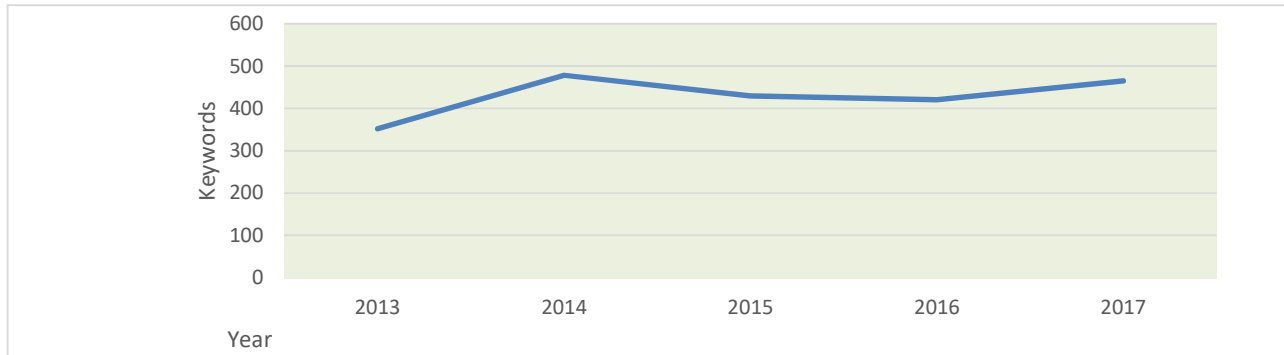


Figure 2. Number of selected keywords over five years.

keywords occurred without repetition. These 112 keywords formed no cluster, and the name given to such keywords is “un-clustered keywords.”. The average frequency of occurrence of un-clustered keywords is 3.4 ($382/112 = 3.4$). The remaining 259 keywords occurred 523 times and formed 72 clusters. Of these 72 clusters, 35 clusters are small, 18 clusters are medium clusters, and 19 clusters are Large in size. As large clusters contain a higher number of relevant keywords, they may be reckoned as “Core” descriptors or core ideas of the subject. Similarly, the medium-sized clusters and small-sized clusters may be considered as “Allied” descriptors and “Alien” descriptors, respectively.

Interestingly, the number of articles peaked sharply in 2014 with 147 publications, then began to decline. The pattern of variation in the selected keywords closely mirrored the trend in the number of articles.

The numerical values of four characteristic indicators (CI) for each large and medium cluster have been calculated and are presented in Table 3. These four indicators are: Integrated Visibility Index, Momentary Visibility Index, Stability Index, and Potency Index, defined as follows (Dutta, Majumder, and Sen, 2010):

Integrated Visibility Index, i.e. $v = F/N$;

Momentary Visibility Index, i.e. $m = F/A$;

Potency Index, i.e. $p = \ln(N \cdot F)$ and

Stability Index, i.e. $s = (A/A_{\max}) \cdot 100$;

where,

$N \sim$ No. of keywords in a cluster

$F \sim$ Frequency of occurrence of all keywords of a cluster

$A \sim$ Occupancy of all keywords in a cluster, i.e. Number of appearance of a keyword over the stipulated time span.

$A_{\max} \sim$ Maximum occupancy of all keywords in the said cluster, i.e. number of keywords of a cluster multiplied by number of years in the said time span.

It is to be noted that frequency of occurrence is different from number of appearance of a keyword. Say, a keyword is found in a database in three years, say 2013, 2014 and 2015 with frequencies 12, 14 and 15 (say). Here the total frequency of occurrence of this keyword in three years is $8 + 7 + 12 = 27$, but its number of appearances is 3, i.e. number of appearance or occupancy just counts in how many years it occurred. The maximum occupancy indicates the highest appearance of a keyword within the time span. For instance, since 2013 to 2017, if a keyword is found only in the years 2013, 2015 and 2017, then its occupancy (A) would be 3, while maximum occupancy (A_{max}) would be 15, i.e. the length of the total time span.

Table 3. Characteristic Indicators (CI) for core descriptors

Large Keyword clusters (Core facets)	(No. of keywords in the cluster, N)	(Frequency of occurrence, F)	(Occupancy, A)	[Maximum occupancy, A(max)]	Integrated Visibility Index, $v = F/N$	Momentary Visibility Index, $m = F/A$	Potency Index, $P = \ln(N*F)$	Stability Index, $S = (A/A_{max})*100$
Black hole	104	560	156	520	5.38	3.58	10.97	30
Cosmology	9	42	25	45	4.66	1.68	5.93	55.55
Einstein's theory	7	11	13	35	1.57	0.84	4.34	37.14
Entropy	19	196	38	190	10.31	5.15	8.22	20
Gravity	31	158	85	310	5.09	1.85	8.49	27.41
Quantum physics	35	112	87	350	3.2	1.28	8.27	24.85
Radiation	8	72	19	80	9	3.78	6.35	23.75
Scalar field	7	20	14	70	2.85	1.42	4.94	20
Space-time	20	81	40	100	4.05	2.02	7.39	40
Tunnelling	13	85	34	130	6.53	2.5	7	26.15
Universe	7	37	25	70	5.28	1.48	5.68	35.71

Table 4. Ranking of core descriptors

CIs Core Facets' Rank	Ranking of core facets in accordance with values of the four CIs			
	Integrated Visibility Index (v)	Momentary Visibility Index (m)	Potency Index (p)	Stability Index (s)
1	Entropy	Entropy	Black hole	Cosmology
2	Radiation	Radiation	Gravity	Space-time
3	Tunnelling	Black hole	Quantum physics	Einstein's theory
4	Black hole	Tunnelling	Entropy	Universe
5	Universe	Space-time	Space-time	Black hole
6	Gravity	Gravity	Tunnelling	Gravity
7	Cosmology	Cosmology	Radiation	Tunnelling

8	Space-time	Universe	Cosmology	Quantum physics
9	Quantum physics	Scalar field	Universe	Radiation
10	Scalar field	Quantum physics	Scalar field	Entropy
11	Einstein's theory	Einstein's theory	Einstein's theory	Scalar field

Table 5. Characteristic Indicators (CI) for allied descriptors

Medium Keyword clusters (Allied facets)	(No. of keywords in the cluster, N)	(Frequency of occurrence, F)	(Occupancy, A)	[Maximum occupancy, A(max)]	Integrated Visibility Index, $v = F/N$	Momentary Visibility Index, $m = F/A$	Potency Index, $P = \ln(N \cdot F)$	Stability Index, $S = (A/A(\max)) \cdot 100$
Anomaly	4	79	16	20	19.75	4.93	5.75	80
Area spectrum	3	21	9	12	7	2.33	4.14	75
Brane theory	4	40	15	20	10	2.66	5.07	75
Conservation theory	5	6	4	25	1.2	1.5	3.4	16
Co-ordinate transformation	4	9	7	25	2.25	1.28	3.58	28
Dirac's theory	5	72	14	25	14.4	5.14	5.88	56
Electromagnetic field	4	6	7	20	1.5	0.85	3.17	35
Energy	6	41	17	30	6.83	2.41	5.5	56.66
Gauge theory	5	37	13	25	7.4	2.84	5.22	52
Horizon	7	84	11	35	12	7.63	6.37	31.42
Information	6	33	16	30	5.5	2.06	5.28	53.33
Particle physics	6	120	22	30	20	5.45	6.57	73.33
Perturbation	4	21	5	20	5.25	4.2	4.43	25
Planck's distribution	3	12	9	15	4	1.33	3.58	60
Quark	4	3	4	20	0.75	0.75	2.48	20
Quasi function	3	27	9	15	9	3	4.39	60
Relativity	3	15	10	15	5	1.5	3.8	66.66
Semiclassical approach	4	9	8	20	2.25	1.13	3.58	40
Space	5	50	14	25	10	3.57	5.52	56

Statistical approach	3	13	8	15	4.33	1.625	3.66	53.33
String theory	3	39	13	15	13	3	4.76	86.66
Thermodynamics	5	90	11	25	18	8.18	6.11	44
Uncertainty principle	5	30	12	25	6	2.5	5.01	48
Wave	4	9	7	20	2.25	1.28	3.58	35
WKB approximation	3	9	8	60	3	1.12	3.29	13.33

Table 6. Ranking of allied descriptors

Allied Facets' Rank	Ranking of allied facets in accordance with values of the four CIs			
	Integrated Visibility Index (v)	Momentary Visibility Index (m)	Potency Index (p)	Stability Index (s)
1	Particle physics	Thermodynamics	Particle physics	String theory
2	Thermodynamics	Horizon	Horizon	Brane theory
3	Dirac's theory	Particle physics	Thermodynamics	Particle physics
4	String theory	Dirac's theory	Dirac's theory	Relativity
5	Horizon	Perturbation	Space	Quasi function
6	Brane theory	Space	Energy	Planck's distribution
7	Space	String theory	Information	Energy
8	Quasi function	Quasi function	Gauge theory	Dirac's theory
9	Gauge theory	Gauge theory	Brane theory	Space
10	Energy	Brane theory	Uncertainty principle	Information
11	Uncertainty principle	Uncertainty principle	String theory	Statistical approach
12	Information	Energy	Perturbation	Gauge theory
13	Perturbation	Information	Quasi function	Uncertainty principle
14	Relativity	Statistical approach	Relativity	Thermodynamics
15	Statistical approach	Relativity	Statistical approach	Semiclassical approach
16	Planck's distribution	Conservation theory	Planck's distribution	Wave
17	WKB approximation	Planck's distribution	Co-ordinate transformation	Electromagnetic field
18	Co-ordinate transformation	Co-ordinate transformation	Wave	Horizon
19	Semiclassical approach	Wave	Semiclassical approach	Co-ordinate transformation
20	Wave	Semiclassical	Conservation	Perturbation

		approach	theory	
21	Electromagnetic field	WKB approximation	WKB approximation	Quark
22	Conservation theory	Electromagnetic field	Electromagnetic field	Conservation theory
23	Quark	Quark	Quark	WKB approximation

Table 7. Rank Correlation Coefficients between different CIs for core and allied descriptors

Rank Correlation Coefficients (r) between the CIs	Values of 'r' for Core Clusters	Nature of correlation	Values of 'r' for Allied Clusters	Nature of correlation
rvm	0.864	Strong +ve	0.867	Strong +ve
rvp	0.391	Weak +ve	0.791	Strong +ve
rvs	0.627	Strong +ve	0.623	Strong +ve
rmp	0.645	Strong +ve	0.892	Strong +ve
rms	0.236	Weak +ve	0.529	Strong +ve
rps	-0.173	Weak -ve	0.729	Strong +ve

Nature of correlation: $1 \geq r \geq 0.5$ indicates Strong +ve Correlation;

$0.5 > r > 0$ indicates Weak +ve Correlation;

$r = 0$ indicates No Correlation;

$0 > r \geq -0.5$ indicates Weak -ve Correlation;

$-0.5 > r \geq -1$ indicates Strong -ve Correlation;

5. Findings

Four crucial feature indicators for the fundamental and related aspects of research keywords are identified in this study: The average frequency of recurrence for each term is the Integrated Visibility Index (v). Momentary Visibility Index (m): The average annual frequency of occurrence for each term. Potency Index (p): A measure of aspect strength that is the natural logarithm of the product of all keywords and their frequency. Stability Index: A measure of temporal stability based on the ratio of actual to maximum occupancy over time. Eleven core, twenty-three

allied, and forty-eight alien descriptions were identified by the research. These indicators' rankings reveal: Entropy, cosmology, black holes, radiation, and gravity are examples of fundamental descriptors. Potency and stability scores for core descriptors are almost inversely correlated, indicating that powerful features are less stable over time. Consistent facet behavior is shown by the significant positive connection that allied descriptors exhibit across all ranks. These relationships were validated using Spearman's Rank Correlation Coefficient, which highlighted distinctions between core and related features.

6. Conclusion:

A technique for measuring the features of a subject's facets has been suggested in this research. Since the facets' behavioral features are quantified using numerical values. Different aspect attributes are implied by the Characteristic Indicators, which in turn describe the subject's particular behavior. For example, it is discovered that the facets created by temporally stable descriptors for "Black Holes" include less low-frequency terms, which may represent a unique characteristic of this topic. But after using this approach on other issues, this would become evident. This research might be expanded to include other branches of the physical sciences as well as more general topics including biology, medicine, agriculture, social science, engineering, cognitive science, behavioral science, etc.

References

1. Berelson, B. 1952. *Content Analysis in Communications Research*. New York: Free Press.
2. Black, J.D. 1962. The Keyword: It's use in abstracting, indexing and retrieving information. *Aslib Proceedings*. 14(10): 313–321.
3. Cambrosio, A., et al. 1993. Historical scientometrics? Mapping over 70 years of biological safety research with Co-word analysis. *Scientometrics*. 27(2): 119–143.
4. Coulter, N., Monarch, I. & Konda, S. 1998. Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science*. 49(13): 1206–1223.
5. Courtial, J.P. 1994. (A) Co-word analysis of scientometrics. *Scientometrics*. 31(3): 251–260.
6. Courtial, J.P., Cahlik, T. & Callon, M. 1994. (A) Model for social interaction between cognition and action through a key-word simulation of knowledge growth. *Scientometrics*. 31(2): 173–192.
7. Dutta, B. 2008. An analytical investigation into the characteristics of a subject - A case study. PhD thesis. Jadavpur University.
8. Dutta, B., Majumder, K.P. & Sen, B.K. 2008. Classification of keywords extracted from research articles published in science journals. *Annals of Library and Information Studies*. 55(4): 317–333.
9. Dutta, B., Majumder, K.P., Sen, B.K. 2010. Content management by keywords: an analytical study. *SRELS Journal of Information Management*. 47(6): 599–620.
10. Hartley, J. & Kostoff, R.N. 2003. How useful are 'keywords' in scientific journals? *Journal of Information Science*. 29(5): 433–438.
11. Index Term. Retrieved on 16 Jan 2016 from https://en.wikipedia.org/wiki/Index_term

12. Retrieved on 5 Mar 2015 from http://shodhganga.inflibnet.ac.in/bitstream/10603/5109/10/10_chapter%201.pdf
13. Law, J. & Whittaker, J. 1992. Mapping acidification research: A test of the co-word method. *Scientometrics*. 23(3): 417–461.
14. Looze, M.D. & Lemarie, J. 1997. Corpus relevance through co-word analysis: An application to plant proteins. *Scientometrics*. 39(3): 267–280.
15. Luhn, H.P. 1960. Keyword-in-Context for technical literature (KWIC Index). *American Documentation*. 11.
16. Noyons, E.C.M. & van Raan, A.F.J. 1998. Advanced mapping of science and technology. *Scientometrics*. 41(1–2): 61–67.
17. Noyons, E.C.M. & van Raan, A.F.J. 1998. Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American Society for Information Science*. 49(1): 68–81.
18. de Solla Price, D.J. 1963. *Little science, big science*. New York: Columbia University Press.
19. de Solla Price, D.J. 1975. *Science since Babylon*. Enlarged ed. New Haven: Yale University Press; p. 165–186.
20. Seetharama, S. 1997. *Information consolidation and repackaging:*

framework, methodology, planning. New Delhi: Ess Ess Pub.