



Scientific measures and tools for research literature output

R. Karpagam, S. Gopalakrishnan¹ and M. Natarajan²

University Library, Anna University, Chennai-600 025, India

¹University Library, MIT Campus, Anna University, Chennai-600 044, India

²Editor & Publisher, Puthiya Parvai, Tamil Arasi Publications, Chennai-600 018, India.

karpagam.au@gmail.com; gopallong@gmail.com; mnindias@yahoo.com

Abstract

The scientometrics research uses various online database, indices and tools in order to establish relationships between authors or their work. This paper describes the scope and limitations of scientometrics - online database, scientometrics indices to perform qualitative/quantitative evaluations and scientometrics tools used to assess the quality of research. The data retrieved from Web of Science on the topic Nanoscience and Nanotechnology during the 2006-2010 i.e. 5 years of records of Indian contributions were analyzed by using various indices.

Keywords: Scientometrics, h-index, g-index, p-index, bibliometrics

Introduction

Scientometrics is the most interesting subject area in the field of library and information science, which can be applied to any discipline irrespective of their period of evolution. It involves quantitative studies of scientific activities, including, among others, publication, and so overlaps bibliometrics to some extent (Tague-Sutcliffe, 1992). Vinkler (2010) stated that scientometrics cannot be restricted with the scope of a scientific discipline. He broadened the definition as quantitative study of people, groups, matters and phenomena in science and their relationships. Chun-Yang Yin (2011), determined the correlation strength between impact factor (JIF), h-index and EigenfactorTM of chemical engineering (CE) journals and its subsequent relevance in indicating the influence and prestige of the journals. He believe that such combination may even apply for other scientific journals as well and this warrants future studies involving bibliometricians for respective fields.

This paper has been divided into four parts, namely, time-line of development, scientometrics online database, scientometrics indices and scientometrics tools which is considered necessary for the scientometrics study.

Table 1. Time-line

Time-line	Description
Early 19 th Century	Origin of bibliometrics research in areas such as law and psychology
1926-48	Lotka's Law, Zipf Law and Bradford Laws developed
1955	Eugene Garfield first describes the Impact factor
1961	Publication of the Science Citation Index
1960s-1970s	Growth of databases made widespread citation analysis a real possibility
1978	Launch of first dedicated journal "Scientometrics"

Time-line of development (Table 1)

Early 19th century the bibliometrics research acquire law and psychology. Laws like Lotka's Law, Zipf Law and Bradford Laws were developed during the period 1926-48. The database were developed during the period

1960's - 1970's. During this 21st century, the database provides more information along with the citation details. It's an added advantage for the scientometrics researchers.

Research information services are being used by scholars, formally and informally to evaluate the research in an efficient and accurate manner. Since the internet has improved the collection and stipulation of citation, practice and right of access metrics, the dispute lies neither in the technology nor the method, but in built databases that deliver services of value. This becomes clear by evaluation of some examples.

Scientometrics online database

Special bibliographic database sources are Web of Science, SciVerse Scopus, Compendex, PubMed, etc. Few of the databases are discussed below. The data can be retrieved from these databases for scientometric study in different format. Example .csv, Refworks, Endnote, Tag format, etc.

Bibliographical databases such as *Web of Science* called Science Citation Index, (SCI), Social Science Citation Index (SSCI) and Arts & Humanities Citation Index (A&HCI) maintained by the Institute for Scientific Information (ISI) in Philadelphia, USA. Web of Science covers over 10,000 of the impact journals worldwide, including Open Access journal and over 110,000 conference proceedings and also the retrospective coverage in the sciences, social sciences, arts and humanities available to 1900 (Thomson Reuters. Retrieved 20.3.2011 from http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/). *Scopus* is an international database. It's easy, quick and comprehensive to find the information scientists need. Contains 41 million records, 70% with abstracts, nearly 18,000 titles from 5,000 publishers worldwide, includes over 3 million conference papers, offers sophisticated tools to track, analyze and visualize research (Elsevier BV. Retrieved on 31.3.2011 from <http://www.info.sciverse.com/scopus/about>). *Compendex*



database provides international coverage of the literature of the engineering field, including civil and structural engineering, computer and electrical engineering, energy technology, materials science and metallurgy, bioengineering, air and water pollution, chemical engineering, and solid waste and hazardous waste management. Citations are drawn from 2,600 journals, technical reports, and conference papers and proceedings. *PubMed* is a free resource, comprises over 20 million citations for biomedical literature from MEDLINE, life science journals, and online books. PubMed citations and abstracts include the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and preclinical sciences. PubMed also provides access to additional relevant Web sites and links to the other NCBI molecular biology resources (National Centre for Biotechnology (NCBI). Retrieved 15.3.2011 from <http://www.ncbi.nlm.nih.gov/books/NBK3827/>).

Scientometrics indices

Standard bibliometric indicators such as number of publications (P) during the study period, number of citations (C) during the study period and the average citation per paper (CPP) have a number of disadvantages. Based on science managers and policy makers request and to support research decisions it is required to increase the bibliometric studies. Different measures and indices have been developed at this level of analysis. One type of indices, such as the *h*-index and *g*-index, describe the most productive core of the output of a researcher and inform about the number of papers in the core. The *h*-index is supposed to measure the broad impact of an individual scientist and to avoid all the disadvantages. Moreover the online database such as Web of Science, Scopus, Google Scholar provides the *h* index. Other indices, such as the *a*-index and *m*-index, depict the impact of the papers in the core.

Bibliometrics methods are used more and more often for evaluation purposes (Pritchard, 1969). Large electronic data bases enable a reasonably fast determination of publication lists and corresponding citation records. But for a comparison of different datasets the dangerous idea to quantify the research output by a single number remains fascinating. Simple indicators as the total number of citations to all papers or the average citation frequency have obvious disadvantages like the difficulty to determine all the citation counts with reasonable accuracy or giving undue weight to highly cited review articles, or taking a possibly large number of irrelevant (not or lowly cited) papers into account. This can be avoided by considering only a small number of relevant or significant papers, but this solution raises the question how to determine this core set of significant papers from a given set of publications. Taking a fixed number or a certain percentage of all publications into consideration would mean a somewhat arbitrary and biased choice. Hence to solve this problem Hirsh introduced *h*-index. Based on this *h* index various indices are developed for Sci. Technol. Edu.

evaluating the career of individual scientists according to their scientific output. Some of the scientometrics measures and indices are discussed in the Table 2.

Table 2. *H index and impact measures*

Index	Introducer	Year	Definition/Formula
<i>h</i> index	Hirsch	2005	A scientist has index <i>h</i> if <i>h</i> of his/her <i>N_p</i> papers have at least <i>h</i> citations each, and the other (<i>N_p</i> - <i>h</i>) papers have no more than <i>h</i> citations each.
<i>g</i> index	Leo Egghe	2006	The highest number <i>g</i> of papers that together received <i>g</i> ² or more citations. From this definition it is already clear that <i>g</i> ≥ <i>h</i>
<i>A</i> index	Jin .B.	2006	$A = \frac{1}{h} \sum cit_j$
<i>h</i> ⁽²⁾ index	Kosmulski	2006	A scientist's <i>h</i> ⁽²⁾ -index is defined as the highest natural number such that his <i>h</i> ⁽²⁾ most cited papers received each at least [<i>h</i> ⁽²⁾] ² citations
Normalized <i>h</i> index (<i>h</i> _{nom})	Sidiropoulos, Katsaros, and Manolopoulos	2007	$h^{*n} = \frac{h}{N_p}$
<i>R</i> index	Jin, B., Liang, L., Rousseau, R., & Egghe, L.	2007	$R = \sqrt{\sum cit_j}$
<i>AR</i> index	Liang et al.	2007	$AR = \sqrt{\sum \frac{cit_j}{a_j}}$
<i>h_w</i> Index	Egghe and Rousseau	2007	$h_w = \sqrt{\sum_{j=1}^{r_0} cit_j}$
<i>e</i> index	Zhang	2009	$e = \sqrt{\sum cit_j - h^2}$
<i>hg</i> index	Alonso	2010	$hg = \sqrt{h \times g}$
<i>p</i> index	Gangan Prathap	2010	$p = h_m = \left(\frac{C^2}{P}\right)^{\frac{1}{3}}$

The *h*-index, also known as the Hirsch index, was introduced (Hirsch, 2005), as an *indicator for lifetime achievement*. Considering a scientist's list of publications, ranked according to the number of citations received, the *h*-index is defined as the highest rank such that the first *h* publications received each at least *h* citations. The *h*-index is not an average, not a percentile, not a fraction; it is a totally new way of measuring performance impact, visibility, quality, etc. of the career of a scientist It is a simple measure without any threshold.

The *g*-index (Egghe, 2006) is an *h*-type index for *quantifying the scientific productivity of physicists and other scientists based on their publication record*. The



index is calculated based on the distribution of citations received by a given researcher's publications. Egghe's *g-index* is rather different from both *h* and h^2 in that it switches attention from the number of most productive papers to the actual number of citations attracted by these most productive papers

A-index (Jin, 2006) achieves the same goal as the *g-index*, namely correcting for the fact that the original *h-index* does not take the exact number of citations of articles included in the *h-core* into account. This index is simply defined as the *average number of citations received by the publications* included in the Hirsch core. The name of this index is derived from the fact that it is just an average (*A*).

The $h^{(2)}$ index (Kosmulski, 2006) which is a *h-type* index that is easier to calculate than the *h-index* since one needs a *shorter list of papers* in decreasing order of number of citations, an author has Kosmulski's index $h^{(2)}$ if $r = h^{(2)}$ is the highest rank such that all papers on ranks $1, \dots, h^{(2)}$ have at least $(h^{(2)})^2$ citations. The number h^2 smaller than the *h* indices and h^2 does not discriminate very well between authors. $g \geq h \geq h^2$ this means that, in principle, accurate determination of the *g-index* requires more work than does the *h-index*, which in turn requires more work than the h^2 -index.

Since *scientists do not publish the same number of articles* (Sidiropoulos *et al.*, 2007), the original *h-index* is not a fair enough metric. Thus, they defined the *Normalized h index* (h_{nom}).

R index is calculated as $R = \sqrt{A \times h}$. In general one way write $R(X, Y)$, where *X* denotes a particular scientist and *Y* the year for which the *R-index* has been calculated. As this is of no importance in our investigations we omit the symbols *X* and *Y*. It is clear that $h \leq R$ as each c_{ij} is at least equal to *h*. In the special case where each c_{ij} is exactly equal to *h*, $R = h$. This nice result is another *advantage of using the square root of the sum*, and not the sum itself.

Liang *et al.* (2007) suggested an *age dependent indicator*: The *AR index* is defined as if a_j denotes the age of article *j* we define the age-dependent *R-index*, denoted by *AR*, by the following equation. If there are several publications with exactly *h* citations then we include the most recent ones in the *h-core*.

Egghe and Rousseau (2008) presented a new *h-index* variation that they called *citation-weighted h-index* (h_w -index) which is, as the *AR-index*, *sensitive to performance changes*. It is clear that indicators that are sensitive to performance changes can be useful in certain environments.

The *e-index* is a necessary *h-index* complement, especially for evaluating highly cited scientists or for *precisely comparing the scientific output of a group of scientists having an identical h-index*. The *e-index* is defined as the square root of the excess citations over

those used for calculating the *h-index* (Zhang, 2009). That is, $e^2 = S(h) - h^2$, where $S(h)$ is the total citations received by the *h* papers for a researcher, if his or her *h-index* is *h*.

hg index is based on a combination of *h-index* and *g-index*, the *hg-index* (Alonso *et al.*, 2010) was proposed as the geometric mean of the *h-index* and the *g-index*. Alonso *et al.* (2010) presented a new index, called *hg-index*, which is based on *both h-index and g-index* that tries to keep a *balance between the advantages of both measures as well as to minimize their disadvantages*. The *hg-index* of a researcher is computed as the geometric mean of his *h-* and *g-* indices, that is: $hg = \sqrt{h \cdot g}$. It is trivial to demonstrate that $h \leq hg \leq g$ and that $hg - h \leq g - hg$, that is, the *hg-index* corresponds to a value nearer to *h* than to *g*. This property can be seen as a penalization of the *g-index* in the cases of a very low *h-index*, thus avoiding the problem of the big influence that a very successful paper can introduce in the *g-index*.

Gangan Prathap (2011) proposed an index called *p-index* (a composite performance index that can effectively combine size and quality of scientific papers) can be extended for scientometrics research assessment in cases where *multiple authorship* is taken into account. The *p-index* strikes the best balance between activity (total citations *C*) and excellence (mean citation rate *C/P*).

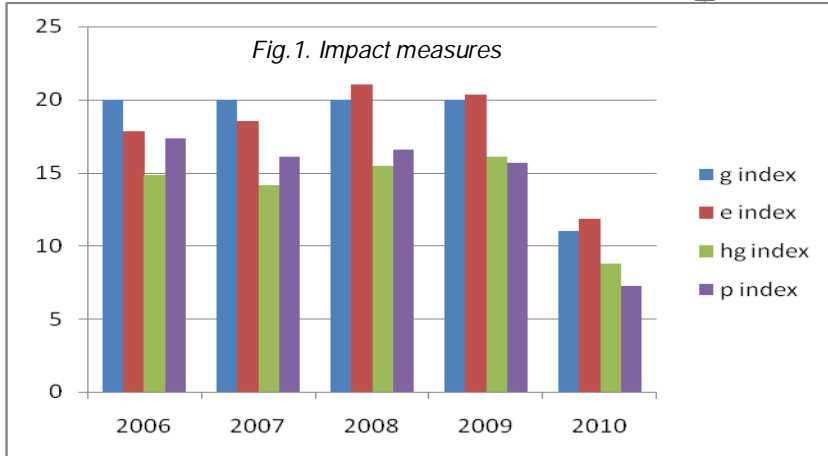
Table 3. Indian contributions on Nanoscience and nanotechnology (2006-2010)

Year	2006	2007	2008	2009	2010	Total
No. of Records	37	47	71	88	95	338
No. of Citations	439	443	569	582	190	2223
Average Citation Per Record	11.86	9.43	8.01	6.61	2	6.58
<i>h</i> index	11	10	12	13	7	23
<i>g</i> index	20	20	20	20	11	36
<i>A</i> index	39.10	44.30	47.42	44.77	27.14	96.65
$h^{(2)}$ index	5	5	5	4	3	7
Normalized <i>h</i> index (h_{nom})	0.30	0.21	0.17	0.15	0.07	0.07
<i>R</i> index	20.95	21.05	23.85	24.12	13.78	47.15
<i>AR</i> index	87.8	110.75	189.67	291	190	444.60
h_w Index	19.39	18.52	18.89	18.71	11.36	32.51
<i>e</i> index	17.83	18.52	21.02	20.32	11.87	41.16
<i>hg</i> index	14.83	14.14	15.49	16.12	8.77	28.77
<i>p</i> index	17.33	16.10	16.58	15.67	7.24	24.45

The *p-index* gives the *best balance between quality (C/P) and quantity (C)*.

Analysis of Indian contribution on nanoscience and nanotechnology research (2006-2010)

A total 338 articles retrieved from Web of Science on the topic Nanoscience and Nanotechnology during the 2006-2010 i.e. 5 years of Indian contributions were analyzed by using the various indices (Table 3).



Based on the 338 total number of publication and 2223 total citations for the period 2006 to 2010, the calculations were made by using the formulas as mentioned in Table 1. While comparing the indices it is observed that $AR > A > R$ and $h > h^2 > h_{nom}$. When compare g index, e-index, p-index and hg-index it is observed that for the year 2006 and 2007 it was $g > e > p > hg$ and for the year 2008, 2009 and 2010 it is *vice versa* and it was shown in Fig.1.

Table 4. Pearson correlations between indices

Indices	ACPP	h	g	A	$h^{(2)}$	h_{nom}	R	AR	h_w	e	hg	p
ACPP	1.000											
h	0.561	1.000										
g	0.849	0.874	1.000									
A	0.627	0.876	0.928	1.000								
$h^{(2)}$	0.923	0.583	0.875	0.782	1.000							
h_{nom}	0.975	0.438	0.730	0.448	0.829	1.000						
R	0.626	0.974	0.934	0.963	0.708	0.475	1.000					
AR	-0.580	0.337	-0.113	0.106	-0.522	-0.655	0.221	1.000				
h_w	0.880	0.871	0.995	0.899	0.885	0.775	0.920	-0.150	1.000			
e	0.630	0.945	0.934	0.983	0.748	0.467	0.994	0.175	0.917	1.000		
hg	0.730	0.968	0.968	0.932	0.755	0.605	0.985	0.114	0.965	0.971	1.000	
p	0.907	0.838	0.989	0.883	0.916	0.808	0.895	-0.220	0.997	0.897	0.944	1.000

It is observed from the above Table 4 and Fig. 2 that the correlation of AR index shows the negative relationship with all the indices except h, A and R indices. Association of h_w and AR index is also in negative aspect. Association of p and h_w is high (0.997), followed by h_w and g index (0.995), e & R index (0.994), p & g index

(0.989).

Scientometrics tools

The quantitative as well as qualitative analysis of online database for scientometrics study, such as, citation mapping, visualization, bibliographic coupling, co-authorship network, co-word mapping etc. are carried out by using Scientometrics tools. These scientometrics tools, purpose and their URL are shown in Annexure 1.

Authormap tool is used for citation mapping and visualization. It is used in ISI Arts & Humanities Citation Index (AHCI), 1988-1997, about 1.26 million records. *Bibcouple* is a tool for visualization of the bibliographic coupling among authors.

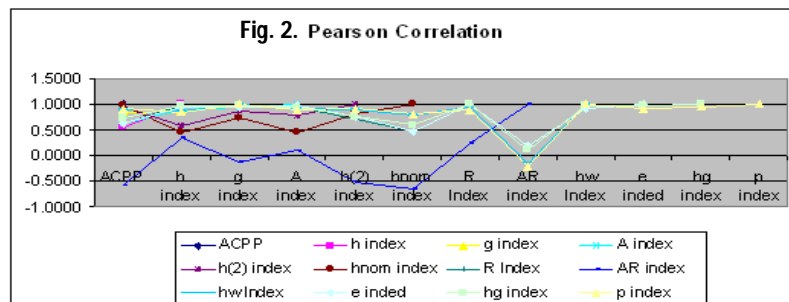
Citespace, is a map type tool mainly for visualizing patterns and trends in scientific literature on citations.

Clean PoP, a web-tool, by Audrey Baneyx/IFRIS, designed to clean systematically the results from the Publish or Perish, another tool in French. *Co-auth* enables to generate a representation of the co-authorship relation in a document set. *Fulltext* software application is useful for co-word mapping of full texts, and also for

word-occurrence matrix. It is available for academic use only.

HistCite is a software application, by Dr. Eugene Garfield, founder of the Institute for Scientific Information and the inventor of the Science Citation Index, which is available on free trial, is useful for various analysis like, complete author list with papers published and citation ranks, complete journal list with papers published and citation ranks, etc. *ISI* is a software application mainly used for converting the bibliographic records downloaded from Web of Science into relational database management.

Patent picture is a commercial tool for analyzing the patents. "Aureka Themescape", proprietary software, renders a patent





Annexure 1. Details of Scientometrics tool

Tool	Purpose	URL	Type	Source	Status	Compatibility
Authormap	Citation Mapping and Visualization	http://project.cis.drexel.edu/authorklink/	Web-tool	Howard White, et. Al. Drexel University, hdwhite@drexel.edu	Other	N/A, FlashPlayer required
Bibcouple	Visualization of the bibliographic coupling among authors using WoS set	http://users.fmg.uva.nl/lleydesdorff/software/bibcoupl/index.htm	Software Application	Loet Leydesdorff	Academic Use Only	PC / DOS. Works with Pajek, MS Access and Excel
Citespace	Visualizing patterns and trends in scientific literature	http://cluster.cis.drexel.edu/%7Ecchen/citespace/	Map	Chaomei Chen,	Other	Images: N/A application: Java required
CleanPoP	Tool is designed to clean results systematically. Publish Or Perish tool	http://cleanpop.ifris.org/guide.html	Web-Tool	Audrey Baneyx/ IFRIS	Public	better with firefox 3 / use browser that respect W3C
Co-auth	Program for visualization of the coauthorship network using a WoS set	http://users.fmg.uva.nl/lleydesdorff/software/coauth/index.htm	Software Application	Loet Leydesdorff	Academic Use Only	PC / DOS. Works with Pajek and MS Access and Excel
Fulltext	Software for co-word mapping of full texts	http://users.fmg.uva.nl/lleydesdorff/software/fulltext/index.htm	Software Application	Loet Leydesdorff	Academic Use Only	PC / DOS. Works with Pajek and MS Excel
HistCite	Bibliographic Analysis and Visualization Software	http://www.histcite.com/index.htm	Software Application	Dr Eugene Garfield, founder of the Institute for Scientific Information and the inventor of the Science Citation Index	Free Trial	PC
IntColl	For Visualization of international collaboration	http://users.fmg.uva.nl/lleydesdorff/software/intcoll/index.htm	Software Application	Loet Leydesdorff	Academic Use Only	PC / DOS. Works with Pajek and MS Access and Excel
ISI	For organizing a set downloaded from the Web-of-Science into databases for relational database management	http://users.fmg.uva.nl/lleydesdorff/software/isi/index.htm	Software Application	Loet Leydesdorff	Academic Use Only	PC / DOS. Works with Pajek and MS Access and Excel
Patent Pictures	It's patently good news	http://www.researchinformation.info/rijanfeb04patents.html	Software Application	Research Information	Commercial	N/A
Publish or perish	Retrieves and analyzes academic citations from Google Scholar	http://www.harzing.com/pop.htm	Software Application	Google Scholar	Freeware	N/A
RefViz	Data visualization and analysis software from the makers of EndNote, ProCite, and Reference Manager for exploring reference collections based on content	http://www.refviz.com/	Software Application	Thomson ResearchSoft	Free Trial	Mac and PC. Interface with EndNote, ProCite, Reference Manager
TI	Co-word mapping of texts	http://users.fmg.uva.nl/lleydesdorff/software/ti/index.htm	Software Application	Loet Leydesdorff	Academic Use Only	PC / DOS. Works with Pajek and MS Excel

landscape, it can plot points based on other characteristics of a patent. *IntColl* is a software application which is used for academic use for visualization of international collaboration.

Publish or Perish (PoP) is a software program that retrieves and analyses citations. It uses Google Scholar to obtain the raw citations. Publish or Perish calculates the citation metrics such as total number of papers, total number of citations, average number of citations per

paper, average number of citations per author, average number of papers per author, Hirsch's h-index and related parameters, Zhang's e-index, Egghe's g-index, the contemporary h-index, etc. *Ref Viz* is for the purpose of the software program is to data visualization and analysis software from the makers of EndNote, ProCite and Reference Manager for exploring reference collections based on content. *TI*, freely available for academic usage, generates a word-occurrence matrix, a word co-



occurrence matrix and a normalized co-occurrence matrix from a set of lines and a word list.

Conclusions

In general, scientometrics analysis use data on numbers and authors of scientific publications and on articles and the citations therein to measure the output of countries, to identify national and international networks, and to map the development of new (multi-disciplinary) fields of science and technology, as well as to know the inner logic of science development. In this paper various indices were discussed. In recent years, the h-index, a measure of the scientific output of researchers based on both the quantity and impact of publications, has received great attention from the scientific community. It uses to measure in order to obtain a more balanced view of the scientific production of researchers and that minimizes some of the problems that they present. Many papers have dealt with this index and have proposed new variations of the h-index to overcome its limitations.

Various indices are well-designed for the scientometrics study. For instance, the h-indexes may increase if in a specific journal of middling or rather low level, groups of researchers intentionally start citing overly each other's work. Just one specific measure is not shrewd to power the assessment of researchers or of research groups. It will strengthen the opinion of administrators and politicians that scientific performance can be expressed simply by one note. Hence, it is suggested that a reliable set of several indicators is necessary, in order to explicate different aspects of performance.

References

1. Alonso S, *et al.* (2009) h-Index: A review focused in its variants, computation and standardization for different scientific fields. *J. Informetrics*. doi:10.1016/j.joi.2009.04.001.
2. Alonso S, Cabrerizo F, Herrera-Viedma E and Herrera F (2010) hg-index: a new index to characterize the scientific output of researchers based on the h- and g-indices. *Scientometrics*. 82, 391-400. DOI 10.1007/s11192-009-0047-5.
3. Chun-Yang Yin (2011) Do impact factor, h-index and Eigenfactor™ of chemical engineering journals correlate well with each other and indicate the journals' influence and prestige? *Curr. Sci.* 100 (5), 648-653.
4. Egghe L and Rousseau R (2008) An h-index weighted by citation impact. *Information Processing & Management*, 44(2), 770-780.
5. Gangan Prathap (2011) The fractional and harmonic p-indices for multiple authorship. *Scientometrics*. 86, 239-244. DOI 10.1007/s11192-010-0257-x.
6. Hirsch JE (2005) An index to quantify an individual's scientific research output (available at http://arxiv.org/PS_cache/physics/pdf/0508/0508-25v5.pdf).
7. Jin (2006) H-index: an evaluation indicator proposed by scientist. *Sci. Focus*. 1(1), 8-9.
8. Jin B, Liang L, Rousseau R and Egghe L (2007) The R- and AR-indices: Complementing the h-index. *Chinese Sci. Bull.* 52(6), 855-863.
9. Kosmulski M (2006) A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter*. 2(3), 4-6.
10. Liang BJL, Rousseau R and Egghe L (2007) The R- and AR-indices: Complementing the h-index. *Chinese Sci. Bull.* 52(6), 855-863.
11. National Centre for Biotechnology (NCBI). Retrieved 15.3.2011 from <http://www.ncbi.nlm.nih.gov/books/NBK3827/>
12. Pritchard A (1969) Statistical bibliography or bibliometrics. *J. Document*. 24(4), 348-349.
13. Sidiropoulos A, Katsaros D and Manolopoulos Y (2007) Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*. 72 (2), 253-280.
14. Tague-Sutcliffe JM (1992) An introduction to informetrics. *Information Processing & Management*. 28, 1-3.
15. Vinkler P (2010) Indicators are the essence of scientometrics and bibliometrics. *Scientometrics*. doi:10.1007/s11192-010-0159-y.
16. Zhang C-T (2009) The e-index, complementing the h-index for excess citations. *PLoS ONE*, 4(5), e5429. doi:10.1371/journal.pone.0005429.