International Journal of Scientometrics, Informetrics and Bibliometrics
ISSN 1137-5019

VOLUME 12 (2008): ISSUE 1. PAPER 2

The h-index of a conglomerate

Ronald Rousseau[1,2], Raf Guns[3], Yuxian Liu[3,4]

[1]KHBO (Association K.U.Leuven), Industrial Sciences and Technology
Zeedijk 101, B-8400 Oostende, Belgium
E-mail: **ronald.rousseau@khbo.be**

[2]K.U.Leuven, Steunpunt O&O Indicatoren & Dept. MSI,
Dekenstraat 2, B-3000 Leuven, Belgium

[3]University of Antwerp, IBW,
Venusstraat 35, B-2000 Antwerpen, Belgium
E-mail: **raf.guns@ua.ac.be**

[4]Periodical Division, Reading Section, Library of Tongji University,
Siping Street 1239, 200092 Shanghai, PR China
E-mail: **yxliu@lib.tongji.edu.cn**

Abstract

Conglomerates are generalized frameworks for informetric research. In this article, the h-index of a conglomerate is defined and it is shown how this construction generalizes the standard definition of the h-index. It is further shown how non-trivial constructions, such as Prathap's h-indices, fit well into a conglomerate framework. An example illustrates the use of the conglomerate framework.

Keywords

h-index; conglomerates; Prathap's institutional h-indices, pseudo h-index

The conglomerate framework (Rousseau, 2005)

If a scientist's publications are ranked in decreasing order of number of citations, then his/her lifetime achievement *h*-index is the highest rank such that the first *h* publications each received *h* or more citations. It is well-known that the idea of an *h*-index can be applied to many source-item relations (Egghe & Rousseau, 2006). Here we present a definition in the general framework of a conglomerate.

A conglomerate, introduced in (Rousseau, 2005), is a framework for informetric (and other) research. Figure 1 illustrates the basic elements of a conglomerate. It consists of two collections and two mappings. The first collection is a finite set, denoted as *S*, and called the source collection. Its elements are called sources. The second collection, denoted as *P*, is called the pool. It is not necessarily finite, but in practical applications it will always be finite. Further a mapping *f* is given from *S* to 2^P, the set of all subsets of *P*. For each $s \in S$, $f(s)$ is a subset of *P*, called the item-set of *s*. The union of all *p* in *P* belonging to at least one item-set is called the item collection, denoted as $I \subset P$. The map *f* itself is called the source-item map.

Each set $f(s)$ is mapped to a number, called the magnitude of this set. This mapping is denoted as *m* and maps $f(s) \in 2^P$ to $m(f(s)) \in \mathbb{R}^+$ (referred to as the *m*-value of source *s*). The mapping itself is called the magnitude function. In

simple cases *m* is the counting measure which maps *f*(*s*) to the number of elements in *f*(*s*). The conglomerate is a quadruple *C* = (*S*,*P*,*f*,*m*).

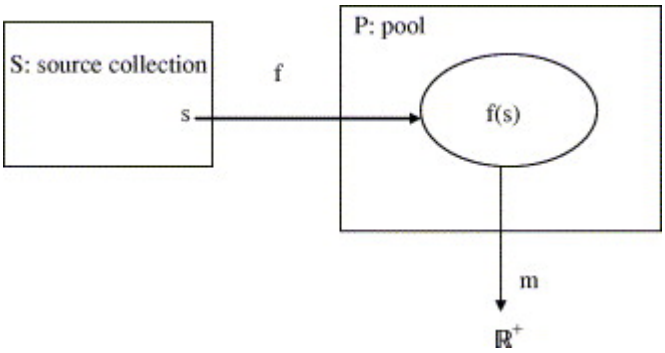

Figure 1. Schematic representation of a conglomerate taken from (Rousseau, 2005)

These steps lead to a first important element in informetric research, namely the ratio of the sum of all magnitudes of item-sets, and the number of elements in the source collection. In a general setting this ratio is referred to as the conglomerate ratio:

$$\text{Conglomerate ratio} = \frac{\sum_{s \in S} m(f(s))}{\#(\text{source collection})}$$

In concrete cases this conglomerate ratio is e.g. a journal's impact factor. Finally, the source-item relation of a conglomerate leads to three lists. The first one just consists of all sources and the magnitude of their corresponding item sets, e.g. articles and corresponding numbers of citations. The second list is the same as the first one, but sources are ranked in decreasing order of the magnitudes of their corresponding item-sets. We will refer to this list as a Zipf list and the rank of a source in this list is called its Zipf rank. The first list can also, if desired and meaningful, be rewritten in size-frequency form (e.g., 7 articles with 1 citation, 4 articles with 2 citations, etc.), leading to a third list associated with the source-item relation of a conglomerate. We may refer to this third list as a Lotka list. Such a list begins with the number of sources that have the lowest magnitude. In case this value is zero such sources are often not mentioned so that the lowest value is usually one (unless fractional counting is applied).

The h-index of a conglomerate

Based on Hirsch's original idea (Hirsch, 2005) we now define the *h*-index of a conglomerate.

Definition

The h-index of conglomerate C is defined as the highest rank such that the magnitude corresponding to Zipf rank *h* is at least equal to *h*. The set of these first *h* sources is called the *h*-core. As usual provisions must be made in case of ties at rank h.

The standard lifetime *h*-index is the *h*-index without any kind of correction for self-citations or co-authorships. Here, the Web of Science (WoS) is used as an example but the procedure is completely equivalent for Scopus or any other citation database. Scientist A's standard lifetime *h*-index, as determined from the WoS, is obtained by taking *S* equal to all articles (co-)authored by A, and collection *P* as all articles included in the WoS. The function *f* maps each article to the set of articles citing this article; *m* is just the counting function, stating how many citations each article has received. Finally, the lifetime *h*-index of scientist A is derived from the corresponding Zipf list. As described here the h-index includes citations to articles outside the group of WoS (or Scopus) journals. By restricting the set S to all articles published in journals covered by the WoS (or Scopus) one obtains the h-index which is usually mentioned. The conglomerate ratio is here equal to the average number of citations received in the WoS (or Scopus) by scientist A.

Most common indicators, including the h-index and the impact factor, can be determined without considering all elements in the pool (the set P), which would be very impractical for large databases. Generally, one only uses the item

collection.

We note that not each conglomerate leads to a meaningful *h*-index, or meaningful conglomerate ratio for that matter. If the lowest magnitude value of the whole source collection is strictly larger than the number of elements in this collection (denoted as N_S), then, by definition this Hirsch index is equal to N_S. Yet, as this may indicate that the h-index construction is not meaningful, we propose to use the lowest value of the corresponding magnitude set in such cases and to refer to it as a pseudo h-value. A pseudo h-value is always larger than or equal to the number of elements in the source collection (which is equal to the h-index in such cases).

Once a Zipf list is obtained it is not difficult to define a g-index (Egghe, 2006) or an R-index (Jin et al., 2007) for a general conglomerate. For an overview of recent developments related to the h-index, including a list of limitations, we refer to (Costas & Bordons, 2007; Rousseau, 2008).

It is rather obvious how existing variations on the h-index idea (e.g. including or not including self-citations; making corrections for co-authorship; making corrections for the position in the byline) can be fitted into the conglomerate framework. Even Eddington's E, defined as the highest number of days in your life on which you have cycled more than E miles, fits easily into this framework. Indeed, the source collection is the set of days in one's life. Each day is mapped to the union of cycling tours made that day (the pool is rather vaguely defined as the set of all cycling tours that can possibly be covered in one day) and the magnitude function maps this union to the distance covered (in miles). The corresponding Zipf list is the ranked list of days in one's life, ranked according to the distance covered during that day. The h-index of this conglomerate is Eddington's E, actually preceding Hirsch' h by more than half a century.

Putting Prathap's approach (Prathap, 2006) in the conglomerate framework is not that easy. For this reason we study this case in more detail.

Prathap's h_1- and h_2–indices presented in a conglomerate framework

In a brief letter published in *Current Science*, Prathap (2006) proposed using two different types of *h*-indices for institutional evaluations: a level one h-index (h_1) and a level two h-index (h_2), where the level one *h*-index is equal to h_1 if the institution (this is the collection of all its researchers) has published h_1 papers, each of which has at least h_1 citations; and its level two h-index is h_2 if the institution has h_2 researchers, each having an individual *h*-index which is at least equal to h_2. If an institute has just one or two high level scientists then the institute's h_1 value will be high but its h_2 will be very low. Another institute that has many scientists of high quality may have approximately the same h_1 but a much higher h_2. In this way the combination of h_1 and h_2 yields useful information about the institute's research structure.

How can h_1 and h_2 be presented in a conglomerate framework? This is not difficult for h_1. As source collection we take the list (ranking plays no role here, so this may be an alphabetical or chronological list) of all articles on which at least one member of the institute has contributed during a given period. This list is denoted as IL (institutional list). As pool P we take any appropriate (local, regional or international) citation database D. The map f_1 maps each article *a* in IL to the set of articles in D citing this article. As usual m_1 is the counting measure, so $m_1(f_1(a))$ is the number of citations received by article *a* in D during a given citation period. We determine the institutional h-index h_1 from the corresponding Zipf list. The conglomerate ratio for this conglomerate is the average number of citations (per article) received by articles written by scientists of this institute.

Describing h_2 in the conglomerate framework is somewhat more complicated. Now we use as source collection the set of all the institute's scientists, denoted as SC. Each scientist (s) is mapped to a set of pairs. The first element of such a pair is an article a(s) written by scientist s, hence an article belonging to IL. The second element of this pair is the set of all articles in D citing article *a(s)*, during the period under study. Hence an image f(s) looks like:

$$\left\{\left(a_{1(s)}, \left\{c_1^{(1)}, c_2^{(1)}, \ldots, c_j^{(1)}\right\}\right), \left(a_{2(s)}, \left\{c_1^{(2)}, c_2^{(2)}, \ldots, c_k^{(2)}\right\}\right), \left(a_{3(s)}, \left\{c_1^{(3)}, \ldots, c_n^{(3)}\right\}\right), \ldots\right\}$$

where $a_{1(s)}, a_{2(s)}$, ... are articles (co)authored by scientist s, and the $c_j^{(y)}$, j=1,2,.... denote the different citing articles of the y^{-th} article written by scientist s. Each *a* belongs to IL, while each set of citing articles belongs to 2^D. The pool corresponding to this situation is IL x 2^D. Hence the source-item map f_2 maps SC to $2^{IL x 2^D}$ (the set of subsets of IL x 2^D). The magnitude function of this conglomerate maps, for each scientist s, the associated set of pairs to this scientist's h-index, denoted as h(s). Observe that each image, f(s), contains exactly the information needed to determine a scientist's h-index. Now the Zipf list associated to this conglomerate is the ranked list of all the h-indices of the institute's scientists. It naturally leads to h_2. The conglomerate ratio of the second conglomerate is the average h-index of all scientists belonging to this institute. In order to clarify this construction we have added a simple fictitious example in the appendix.

This section shows how Prathap's h-indices can be described in a conglomerate framework. As the construction of h_2 is a special case of the construction of successive *h*-indices (Schubert, 2007) this construction also shows how successive *h*-indices can be described in a conglomerate framework.

We note that if this construction is applied to a small research group, then the smallest author h-index can easily be (much) larger than the number of researchers in the group. This would be an example where a pseudo h-index could be given as a meaningful indicator, instead of or besides h_2, which, in this case, would be equal to the number of researchers in the group. An example of a practical calculation of Prathap indices is given in (Arencibia-Jorge & Rousseau, 2009).

Conclusion

Conglomerates form very general frameworks in which many different kinds of informetric (and other) research can be presented. It is shown how the conglomerate framework is also a natural environment for the study of h-indices. This definition leads to a huge generalization of the original concept. Indeed, as is indicated in the text, all kinds of variants of Hirsch' original proposal fit easily into the conglomerate framework by changing one or more of the conglomerate's elements. Even non-trivial extensions, such as Prathap's institutional h-indices, can be presented as conglomerates. It is observed that once the Zipf list of a conglomerate is drawn, other h-type indices such as Egghe's g and the R-index can also be generalized to the conglomerate framework. We finally note that as we do not deal with pure mathematics, but with real-world applications in many fields, such an abstract framework does not always lead to a meaningful result. Moreover, people may have different opinions as to the meaning of the term "meaningful". We leave it to our colleagues to use common sense and apply a "reality check" when applying our ideas.

Acknowledgement

The authors thank two anonymous referees for some interesting suggestions improving the presentation of this article. This work was also supported by the National Natural Science Foundation of China (NSFC Grant No 70673019).

References

Arencibia-Jorge, R. and Rousseau, R. (2009). Influence of individual researchers' visibility on institutional impact: an example of Prathap's approach to successive *h*-indices. *Scientometrics* (2009: to appear).

Costas, R and Bordons, M. (2007). The h-index: advantages, limitations and its relation with other bibliometric indicators at the micro-level. *Journal of Informetrics*, 1(3), 193-203.

Egghe, L. (2006). An improvement of the H-index: the G-index. *ISSI Newsletter*, 2, 8-9.

Egghe, L. and Rousseau, R. (2006). An informetric model for the *h*-index.

Scientometrics, 69(1): 121-129.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences, 102(46): 16569-16572.

Jin, BH., Liang, LM, Rousseau, R. and Egghe, L. (2007). The R- and AR-indices: complementing the h-index. *Chinese Science Bulletin*, 52, 855-863.

Prathap, G. (2006). Hirsch-type indices for ranking institutions' scientific research output. *Current Science*, 91, 1439.

Rousseau, R. (1997). Sitations: an exploratory study. Cybermetrics, vol.1(1): Paper 1. **http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html**

Rousseau, R. (2005). Conglomerates as a general framework for informetric research. *Information Processing and Management*, 41(6): 1360-1368.

Rousseau, R. (2008). Reflections on recent developments of the h-index and h-type indices. *COLLNET Journal of Scientometrics and Information Management*, 2(1), 2008, 1- 8.

Schubert, A. (2007). Successive h-indices. *Scientometrics*, 70, 201-205.

Appendix

A fictitious example: The SMALL-IS-BEAUTIFUL Institute

The SMALL-IS-BEAUTIFUL Institute is a research institute with only four scientists. Their publication-citation record over the investigated period is given in Table 1.

Table 1. Publication-citation record of the (fictitious) SMALL-IS-BEAUTIFUL Institute

SC	IL	Citing articles
SC1	$ART_{SC1}1$	Citingarticle1
		Citingarticle2
		Citingarticle3
	$ART_{SC1}2$	Citingarticle1
		Citingarticle3
	$ART_{SC1}3$	----------
SC2	$ART_{SC2}1$	Citingarticle4
SC3	$ART_{SC3}1$	Citingarticle5
	$ART_{SC3}2$	Citingarticle6
		Citingarticle7
		Citingarticle8
		Citingarticle9
		Citingarticle10
SC4	$ART_{SC4}1$	----------
	$ART_{SC4}2$	Citingarticle6
	$ART_{SC4}3$	Citingarticle6
		Citingarticle7
		Citingarticle8
	$ART_{SC4}4$	Citingarticle6
		Citingarticle7
		Citingarticle8
		Citingarticle11

The second column contains IL and we assume that a certain P has been chosen from which the citing articles have been retrieved.

The first source-item map f_1 : IL -> 2^P is defined as follows:

$ART_{SC1}1$ -> $f_1(ART_{SC1}1)$ ={Citingarticle1, Citingarticle2, Citingarticle3}
$ART_{SC1}2$ -> $f_1(ART_{SC1}2)$ ={Citingarticle1, Citingarticle3}
$ART_{SC1}3$ -> $f_1(ART_{SC1}3)$ = Ø
$ART_{SC2}1$ -> $f_1(ART_{SC2}1)$ ={Citingarticle4}
$ART_{SC3}1$ -> $f_1(ART_{SC3}1)$ ={Citingarticle5}
$ART_{SC3}2$ -> $f_1(ART_{SC3}2)$ ={Citingarticle6, Citingarticle7, Citingarticle8,

Citingarticle9, Citingarticle10}
$ART_{SC4}1$ -> $f_1(ART_{SC4}1) = \emptyset$
$ART_{SC4}2$ -> $f_1(ART_{SC4}2)$ ={Citingarticle6}
$ART_{SC4}3$ -> $f_1(ART_{SC4}3)$ ={Citingarticle6, Citingarticle7, Citingarticle8}
$ART_{SC4}4$ -> $f_1(ART_{SC4}4)$ ={Citingarticle6, Citingarticle7, Citingarticle8, Citingarticle11}

For m_1 we use the counting measure.

Then $m_1(f_1(ART_{SC1}1)) = 3$
$m_1(f_1(ART_{SC1}2)) = 2$
$m_1(f_1(ART_{SC1}3)) = 0$
$m_1(f_1(ART_{SC2}1)) = 1$
$m_1(f_1(ART_{SC3}1)) = 1$
$m_1(f_1(ART_{SC3}2)) = 5$
$m_1(f_1(ART_{SC4}1)) = 0$
$m_1(f_1(ART_{SC4}2)) = 1$
$m_1(f_1(ART_{SC4}3)) = 3$
$m_1(f_1(ART_{SC4}4)) = 4$

This leads to the Zipf list

$ART_{SC3}2$	5
$ART_{SC4}4$	4
$ART_{SC1}1$	3
$ART_{SC4}3$	3
$ART_{SC1}2$	2
$ART_{SC2}1$	1
$ART_{SC3}1$	1
$ART_{SC4}2$	1
$ART_{SC1}3$	0
$ART_{SC4}1$	0

and hence $h_1 = 3$.

The second source-item map f_2 : SC-> $2^{IL \times 2^D}$ is defined as follows:

SC1 -> { ($ART_{SC1}1$, {Citingarticle1, Citingarticle2, Citingarticle3}),
($ART_{SC1}2$, {Citingarticle1, Citingarticle3}),
($ART_{SC1}3$, $\emptyset$)}
SC2 -> { ($ART_{SC2}1$, {Citingarticle4})}
SC3 -> { ($ART_{SC3}1$, {Citingarticle5}),
($ART_{SC3}2$, {Citingarticle6, Citingarticle7, Citingarticle8, Citingarticle9, Citingarticle10})}
SC4 -> { (ARTSC41, $\emptyset$),
($ART_{SC4}2$, {Citingarticle6}),
($ART_{SC4}3$, {Citingarticle6, Citingarticle7, Citingarticle8 }),
($ART_{SC4}4$, {Citingarticle6, Citingarticle7, Citingarticle8, Citingarticle11}) }

The magnitude function m_2 associates the standard h-index to an image of f_2.

Then $m_2(f_2(SC1)) = 2 = h(SC1)$
$m_2(f_2(SC2)) = 1 = h(SC2)$
$m_2(f_2(SC3)) = 1 = h(SC3)$
$m_2(f_2(SC4)) = 2 = h(SC4)$

This leads to the Zipf list

SC1	2
SC4	2
SC2	1
SC3	1

and hence $h_2 = 2$.

As suggested by a referee we also add a Lotka list, namely the one for the first conglomerate (and determined by f_1 and m_1). This list looks as follows:

2 articles with 0 citations
3 articles with 1 citation
1 article with 2 citations

2 articles with 3 citations
1 article with 4 citations
1 article with 5 citations

The first line is often omitted from the list as authors with no publications are usually not mentioned. Rousseau (1997) offers an exception in the case of inlinks (sitations) on the Internet (by including sites with no inlinks and fitting a shifted Lotka function).

Received 28/August/2008
Accepted 10/October/2008