Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing and Management 41 (2005) 1360–1368

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Conglomerates as a general framework for informetric research

Ronald Rousseau

KHBO (Association K.U. Leuven), Industrial Sciences and Technology, Zeedijk 101, B-8400 Oostende, Belgium
University of Antwerp, IBW, Universiteitsplein 1, B-2610 Wilrijk, Belgium

Received 27 September 2004; accepted 3 March 2005
Available online 26 April 2005

Abstract

We introduce conglomerates as a general framework for informetric (and other) research. A conglomerate consists of two collections: a finite source collection and a pool, and two mappings: a source-item map and a magnitude map. The ratio of the sum of all magnitudes of item-sets, and the number of elements in the source collection is called the conglomerate ratio. It is a kind of average, generalizing the notion of an impact factor. The source-item relation of a conglomerate leads to a list of sources ranked according to the magnitude of their corresponding item-sets. This list, called a Zipf list, is the basic ingredient for all considerations related to power laws and Lotkaian or Zipfian informetrics. Examples where this framework applies are: impact factors, including web impact factors, Bradford–Lotka type bibliographies, first-citation studies, word use, diffusion factors, elections and even bestsellers lists.
© 2005 Elsevier Ltd. All rights reserved.

Keywords: Informetric studies; Citation analysis; Impact factors; Power laws; Inlinks; Social networks; Collaboration; Demography; Writers; Diffusion; Elections

1. Introduction

The field of informetrics includes citation analysis, obsolescence studies, word-related analyses, author studies, growth, studies of the so-called informetric laws (Lotka, Zipf, Bradford, Mandelbrot, Leimkuhler), and webometrics (Wilson, 1999). In this article we present a general framework for informetric studies. It is shown that the same framework applies to many situations, also outside the information sciences. A new term 'conglomerate' is introduced and its use in this context explained.

E-mail address: ronald.rousseau@khbo.be

0306-4573/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2005.03.007

We begin by shedding a new light on the ISI impact factor and on Bradford's bibliographies. Then the new framework is introduced and many examples of its use are given, including different types of impact factors, author production and co-authorship data, web impact factors, demographical studies, bestsellers lists, word use, diffusion factors, elections and relative impact factors.

2. A fresh look at the classical impact factor and at Bradford's bibliography on Applied Geophysics

Consider the collection of all 'citable' articles published in journal J during the years $Y - 1$ and $Y - 2$. Consider further the collection of all references in journals covered by the Web of Science in the year Y, and its subcollection consisting of all items referring to articles published in journal J during the years $Y - 1$ or $Y - 2$. We count the number of elements in this subcollection, and the number of elements in the collection of 'citable' articles. Finally taking the ratio of these two numbers yields the classical impact factor of journal J in the year Y (Garfield & Sher, 1963). We may also make a list of all these articles ranked according to the number of citations received.

Somewhat similarly we may consider the collection of all journals that have published at least one article about the subject *Applied Geophysics* in the period 1928–1931. Next we consider the collection of all scientific articles published worldwide during the period 1928–1931 (about any subject) and consider the subcollection of articles dealing with *Applied Geophysics*. We count the number of articles about *Applied Geophysics* and the number of journals containing at least one of these articles. Their ratio is just the average number of articles on *Applied Geophysics* (in journals containing at least one such article). Finally we draw the list of all journals and the corresponding number of articles they contain on *Applied Geophysics*. The size-frequency form of this list can be found in Bradford's original article (Bradford, 1934) and is reproduced in (Egghe & Rousseau, 1990). It was the basis for Bradford's famous law of scattering.

We presented these two classical examples in such a way that correspondences are made clear. Next we introduce the notion of a conglomerate and highlight the steps leading to this general framework.

3. Conglomerates

A conglomerate is a framework for informetric (and other) research. 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 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 will be called the source-item map.

In the previous examples the source collections were all 'citable' articles published in journal J during the years $Y - 1$ and $Y - 2$, and all journals that published at least one article on *Applied Geophysics* in the period 1928–1931. The pool of the first example was the collection of all ordered pairs consisting of a citing article and an item in this article's reference list. These citing articles are all published in journals covered by the Web of Science in the year Y. The items of this reference list are represented in such a way that no citing article can be confused with another citing article and each item from a reference list, published in the period $[Y - 2, Y - 1]$ can unequivocally be traced to an article in the source collection. For the second case the collection of all scientific articles published worldwide during the period 1928–1931 form the pool. Source-item maps were not explicitly given. For the impact factor example the function f maps each article

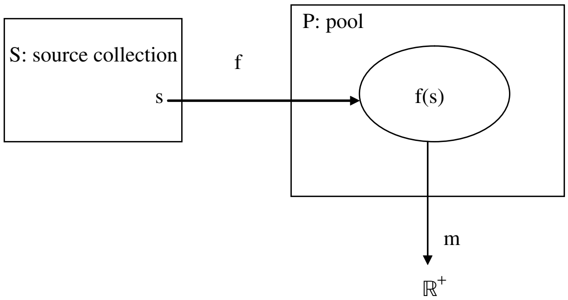

Fig. 1. Schematic representation of a conglomerate.

to the set of all its citations; in the Bradford case the function maps each journal to the set of articles published in this journal and dealing with *Applied Geophysics*.

Fig. 1 illustrates the basic elements of a conglomerate.

Next, 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 $m(f(s)) \in \mathbb{R}^+$. The mapping itself is called the magnitude function. Often, but not always (this is one of the generalizations with respect to the classical theory), m will be the counting measure which maps $f(s)$ to the number of elements in $f(s)$. This was the case for the two examples.

These steps lead to a first important topic in informetric research, namely the ratio of the sum of all magnitudes of item-sets, and the number of elements in the source collection. This ratio will in general be referred to as the conglomerate ratio.

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

In our examples the conglomerate ratios are the classical impact factor and the average number of articles on *Applied Geophysics* in journals containing at least one such article (the average production).

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. In the Bradford case this list consists of all journals that published at least one article on *Applied Geophysics* in the period 1928–1931 and the corresponding number of published articles. The second list is the same as the first one, but sources are ranked according to the magnitudes of their corresponding item-sets. We will refer to this list as a Zipf list. The first list can also, if desired, be rewritten in size-frequency form, leading to a third list associated with the source-item relation of a conglomerate (Egghe, 2005; Rousseau & Rousseau, 1993). These lists are the basic ingredients for all considerations related to power laws and Lotkaian or Zipfian informetrics (Egghe, 2005). The corresponding source-item relation can also be represented by a (weighted) Lorenz curve and measures of inequality such as the Gini index can then be used as a characteristic number for the source-item relation (Burrell, 1991; Egghe, 2005; Rousseau, 1992).

Why do we refer to this framework as a 'conglomerate'? A conglomerate is either a business venture whose asset growth comes through the acquisition of, or merger with, other firms whose products are largely unrelated to each other; or in geology, the term refers to sedimentary rock composed largely of pebbles or other rounded particles. In both cases it refers to a whole composed of a finite number of discrete parts as is the case here for the source collection. In mathematics it is customary to name new structures by giving existing words a new meaning (examples are: group, ring, module, sheaf). It is this tradition that we follow. Note that we have used the term 'conglomerate' before (Egghe & Rousseau, 2005) but there its meaning was more restricted.

What are the differences and similarities between conglomerates as introduced here, and Egghe's information production processes or IPPs (Egghe, 1990)? As for IPPs we have two collections and a mapping from sources to an item-set. The difference is that we restrict the source set to be a finite set. In this sense conglomerates are less general than IPPs. On the other hand we make the pool explicit, and introduce the magnitude function m as part of the framework.

3.1. An observation about relative rankings

In the examples given above, and also those that follow, it is possible to give the values of the magnitude map for each source separately. This is not so for the relative case. In this case the relative magnitude, denoted as m_r is given as:

$$m_r(s) = \frac{m(f(s))}{\sum_{t \in S} m(f(t))}$$

Clearly all m-values must be known before any m_r-value can be calculated. Conglomerate ratios based on relative magnitudes have no importance: they are always equal to the reciprocal of the number of elements in the source collection. The corresponding Zipf list yields the same ranking as the original one, but the values might be more meaningful as in the well known case of the relative impact factor (Braun, Glänzel, & Schubert, 1985; Egghe & Rousseau, 2003; Schubert, Glänzel, & Braun, 1983).

We will now show how, besides the classical impact factor and Bradford-type bibliographies, other informetric topics fit into the conglomerate framework. Note that all cases considered further on are just examples. Much more cases can be put in the conglomerate framework. Moreover, in each example that we will consider changes can be made to the source collection, the pool or the item collection, or to more than one of these sets. All these changes result in other examples of conglomerates. We are convinced that the reader can make these small generalizations for him/herself.

4. Examples of conglomerates

4.1. Impact factors

The Garfield–Sher impact factor is a synchronous impact factor (Ingwersen, Larsen, Rousseau, & Russell, 2001). Other synchronous impact factors are obtained by using a different period for the source collection. Instead of the time period $[Y-2, Y-1]$ a general period $[Y-n-s+1, Y-s]$ is then used. The source-item map sends each article in the source collection to the set of citations received in the year Y. As for the Garfield–Sher impact factor also here the simple counting measure is used. This yields an n-year synchronous impact factor, with offset period s (for the Garfield–Sher impact factor, the offset period is 1 year).

Diachronous impact factors are obtained by considering one fixed year Y for the source collection, and a time period of the form $[Y, Y+n-1]$ for the pool, and hence also for the item collection. The conglomerate ratio for this situation is the n-year diachronous impact factor. Generalized impact factors (Frandsen & Rousseau, 2005) can be described in a similar way.

The role of the pool can be illustrated by considering a journal such as the *Chinese Science Bulletin*. Its ISI impact factor is obtained by using as its pool articles published in journals referenced by the Web of Science. Its CSCD (Chinese Science Citation Database) or CSTPC (Chinese Science and Technology Publications and Citations) impact factor is obtained by using the CSCD or the CSTPC as pool (Jin & Wang, 1999; Wu et al., 2004). The role of the pool is also very important in comparisons of impact factors over different years. For instance, Garfield–Sher impact factors calculated using ISI's retrospective data for the period (1945–1964) are usually smaller than those for later years. One of the reasons for this is the fact that the pool is smaller than that for later years.

4.2. Author production data

In this typical example the source collection consists of first authors of articles dealing with a certain topic *T* written during a period *P*. The pool may consist of all articles present in a database. The source-item map sends each author to the set of articles dealing with a certain topic *T* written during a period *P* and for which the scientist is first author. Using again the counting measure leads to the average number of articles per scientist (as first author) as conglomerate ratio. The corresponding Zipf list ranks authors according to the number of articles first-authored by them (about topic *T* during the period *P*) and included in the database. The corresponding size-frequency list is the type of list drawn by Lotka (1926), leading to the discovery of the power law bearing his name.

What happens if one considers all authors, and not just first authors? In this case the source collection is a group of scientific investigators, say medical scientists. The pool collection could be Medline, and the source-item map sends each scientist to the set of articles (in Medline) for which this scientist is the author or a co-author. Using the counting measure again leads to the average production of this group of scientists (using the normal count procedure) as conglomerate ratio. The Zipf list ranks each scientist according to his/her total production (in Medline).

If, however, the magnitude measure maps each set of articles to the sum of fractional contributions of the corresponding scientist, then this leads to an average production and a Zipf production list according to the fractional counting procedure. This is the advantage of having a general magnitude map in the conglomerate framework, and not only the counting map.

4.3. Collaboration

Alternatively, one may consider as source collection all articles authored or co-authored by a targeted group of scientists. As pool we consider a large group of scientists including the targeted group, and the source-item function maps each article to the set of its authors. Using again the counting measure leads to the average number of authors per paper. Note that the total sum in the numerator for the conglomerate ratio is not equal to the number of scientists but to the total number of (possibly overlapping) contributions each of them made. The corresponding Zipf list ranks articles according to the number of co-authors.

4.4. First citation and most recent reference studies

As source collection we consider a journal issue (or any other collection of articles). The pool is again the ISI database. Each article, s, is mapped to the first article (in the pool) citing it. So, in this case, $f(s)$ is a singleton. The magnitude function m maps $f(s)$ to the time between the publication of s and its citation in $f(s)$. The conglomerate ratio is just the average time between publication and citation. First citation studies (see e.g. Burrell, 2001; Egghe, 2000; Egghe & Rao, 2001; Rousseau, 1994) are studies of the corresponding conglomerate lists. A similar description can be given for most recent reference studies (Egghe & Rao, 2002).

4.5. Networks: web impact factors and directed network density

As source collection we take e.g. all web pages belonging to a fixed top-level domain (e.g. .com, or .uk, .be, .cn, ...). As pool we take the World Wide Web considered as a collection of inlinks from higher level domains to higher level domains. The source-item map sends each webpage to its set of inlinks. Using again the simple counting measure yields the web impact factor as conglomerate ratio (Ingwersen, 1998; Thelwall, 2000). All web pages in the source collection can now be ranked according to their 'popularity', i.e. number of inlinks.

The World Wide Web is a special case of a network. Consider now a general directed network as studied in social network analysis (Wasserman & Faust, 1994). As source collection we consider all ordered pairs of nodes in a given network. The pool consist of all possible arcs in this network, and the source-item mapping send each pair either to the directed arc connecting this pair, or to the empty set if this arc is not present. Using the counting measure yields as conglomerate ratio the number of arcs present divided by the number of ordered pairs of nodes. This is exactly the density of a directed network (Wasserman & Faust, 1994, p. 129). The corresponding Zipf list is less interesting: it consists of two parts, namely the connected node pairs (with magnitude one) and those that are not connected (with magnitude zero). It gives the elements of the corresponding adjacency matrix or sociomatrix.

4.6. Cities and their inhabitants

In demography and geography Zipf's law is well known too (Gabaix, 1999; Ioannides & Overman, 2003; Zipf, 1941). Here the source collection may consist of all cities in a country or region. The pool is just all inhabitants in the world. The source-item map sends each city to its set of inhabitants. Using the counting measure leads to the average number of inhabitants per city in this country or region. Zipf's list ranks cities according to the number of inhabitants.

4.7. Fiction authors and sales

The following example is situated in the domain of publishing. Consider a fiction author such as Kurt Vonnegut. The source collection is the set of all books written by this author. The pool may consist of all book copies sold in the year Y in the USA. The source-item map sends each title to the set of copies sold in the year Y in the USA. In this example the conglomerate ratio is the average number of copies per title sold (again using the counting measure), a number representing the sales impact of Kurt Vonnegut in the year Y in the USA. Using as magnitude measure the profit made by the sales of one title yields another interpretation of the conglomerate ratio: in this case it is the average profit per book title. The corresponding Zipf list (=Vonnegut's personal bestseller list) consists of all book titles ranked according to the number of copies sold (or profits made).

More generally, the source collection may consist of a group of fiction authors, and again the pool may consist of all book copies sold in the year Y in the USA. The source-item map sends each author to the set of copies sold in the year Y in the USA. In this example the conglomerate ratio is the average number of copies sold for this group of writers (again using the counting measure). The corresponding Zipf list is a (yearly) bestseller list per author.

4.8. Word use

This is another classical case. The source collection consists of all different word types used in a document. The pool consists of all word uses (tokens) in this document and the source-item map sends each word type to its set of tokens in this document. The counting measure gives as conglomerate ratio the average number of uses per word type. A typical Zipfian list results from ranking word types according to their use in this document (Zipf, 1949).

4.9. Diffusion

Consider the collection of all citations of articles published in a fixed journal J, cited during the year Y, where citations refer to articles published during the period $[Y - n + 1, Y]$. These citations are ranked in an arbitrary but fixed order (e.g. as indexed in the Web of Science). This leads to the source collection $S = \{s_1, s_2, \ldots, s_m\}$. Consider further the collection of all journals indexed in the Web of Science as pool, and map each citation to the journal in which it was cited: $f(s_j)$ is the journal in which reference s_j was published. The magnitude mapping is constructed as follows: $m(f(s_1)) = 1$, $m(f(s_2)) = 1$ if $f(s_2)$ is not equal to $f(s_1)$ and $m(f(s_2)) = 0$ is $f(s_1) = f(s_2)$. In general: $m(f(s_j)) = 1$ if $f(s_j)$ does not belong to the set $\{f(s_1), \ldots, f(s_{j-1})\}$, and is equal to 0 if $f(s_j)$ belongs to this set. The resulting conglomerate ratio is equal to the number of different journals used in citing the journal J in the year Y divided by the total number of citations received that year. This is nothing but the relative synchronous diffusion factor introduced by Rowlands (2002) (see also Frandsen, Rousseau, & Rowlands, 2005). Note that this diffusion factor is not influenced by the arbitrary order in which we ranked the citations of the source collection. In this approach the Zipf list has no meaning, because it does depend on the arbitrary ranking of the source collection.

Another magnitude mapping, denoted as n, leads to the same relative synchronous diffusion factor but has the advantage that the resulting Zipf list is meaningful. Indeed, $n(f(s_j))$ is defined as 1/(the number of times the journal $f(s_j)$ is an image of the source-item map f). The contribution of one journal to the numerator of the conglomerate ratio is still 1 (as for the m-map), so that we still obtain a relative synchronous diffusion factor. Yet, the Zipf list now ranks each citation to its contribution in the corresponding journal. More precisely if citation s_j is one of the 5 references published in journal J_t then the share of citation s_j to J_ts contribution in the diffusion of journal J is 1/5. This Zipf list is headed by unique citations (unique with respect to the journal in which this citation occurred).

Similarly, one may consider as source the collection of all citations of a fixed journal J occurring during the period $[Y, Y + n - 1]$, where citations refer to articles published during the year Y. Consider again the collection of all journals indexed in the Web of Science as pool. Using again the same source-item mapping f, and magnitude measures m or n leads to the relative diachronous diffusion factor as conglomerate ratio (Frandsen et al., 2005; Rowlands, 2002).

Finally, consider as source collection the set of articles published in the journal J during the year Y, and all journals indexed in the Web of Science as pool. As source-item function we send each article to the set of journals that cite this article during the period $[Y, Y + n - 1]$. If articles are numbered then $m(f(s_j))$ is equal to the number of different journals not belonging to the union of the sets $f(s_1), \ldots, f(s_{j-1})$. The conglomerate ratio is here the ratio of the number of different journals used for citing and the total number of publications. This is the diffusion factor as introduced by Frandsen (Frandsen, 2004; Frandsen et al., 2005).

4.10. Elections

In this example, the source collection consists of all candidates for an election. The pool consists of all voters, and the source-item map sends each candidate to the set of people who voted for them. The magnitude map is just the counting map if each voter has exactly one vote. If each voter may, however, place weights on votes (e.g. 6 points for the best candidate, 3 points to the second best and 1 point for the third best) then the magnitude map is the sum of all points received by a candidate. The conglomerate ratio is not interesting here as it is just the average score of each candidate. In elections it is the ranked Zipf list of scores per candidate that matters.

5. Conclusion

In this article we presented a general framework for bibliometric–scientometric–informetric–webometric–cybermetric studies. One of the advantages of this framework is that fractional counts and other weighting schemes find a natural place in it. Examples where this framework applies are: impact factors, including web impact factors, Bradford–Lotka type bibliographies, word use, diffusion factors and even bestsellers lists and elections. These examples make it clear that the same framework applies to many situations, also outside the information sciences. Studying conglomerates at consecutive moments in time leads to dynamic studies of the field.

Acknowledgment

The author thanks Leo Egghe, editor of this special issue, and Thierry Lafouge for helpful comments on this article.

References

Bradford, S. (1934). Sources of information on specific subjects. *Engineering, 137*, 85–88.

Braun, T., Glänzel, W., & Schubert, A. (1985). *Scientometric indicators. A 32-country comparative evaluation of publishing performance and citation impact*. Singapore: World Scientific.

Burrell, Q. L. (1991). The Bradford distribution and the Gini index. *Scientometrics, 21*, 181–194.

Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics, 52*, 3–12.

Egghe, L. (1990). The duality of informetric systems with applications to the empirical laws. *Journal of Information Science, 16*, 17–27.

Egghe, L. (2000). A heuristic study of the first-citation distribution. *Scientometrics, 48*, 345–359.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Oxford: Elsevier.

Egghe, L., & Rao, I. K. R. (2001). Theory of first-citation distributions and applications. *Mathematical and Computer Modelling, 34*, 81–90.

Egghe, L., & Rao, I. K. R. (2002). Theory and experimentation on the most-recent reference distribution. *Scientometrics, 53*, 371–387.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics*. Amsterdam: Elsevier.

Egghe, L., & Rousseau, R. (2003). A general framework for relative impact indicators. *Canadian Journal of Information and Library Science/Revue canadienne des sciences de l'information et de bibliothéconomie, 27*, 29–48.

Egghe, L., & Rousseau, R. (2005). Comparing partial and truncated conglomerates from a concentration theoretic point of view. *Mathematical and Computer Modelling, 41*, 301–311.

Frandsen, T. F. (2004). Journal diffusion factors—a measure of diffusion? *Aslib Proceedings, 56*, 5–11.

Frandsen, T. F., & Rousseau, R. (2005). Article impact calculated over arbitrary periods. *Journal of the American Society for Information Science and Technology, 56*, 58–62.

Frandsen, T. F., Rousseau, R., & Rowlands, I. (2005). Diffusion factors. Preprint.

Gabaix, X. (1999). Zipf's law and the growth of cities. *American Economic Review, 89*, 129–132.

Garfield, E., & Sher, I. H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation, 14*, 195–201.

Ingwersen, P. (1998). The calculation of Web impact factors. *Journal of Documentation, 54*, 236–243.

Ingwersen, P., Larsen, B., Rousseau, R., & Russell, J. (2001). The publication-citation matrix and its derived quantities. *Chinese Science Bulletin, 46*(6), 524–528, Also published in Chinese: *Kexue Tongbao, 46*(8), 700–704.

Ioannides, Y. M., & Overman, H. G. (2003). Zipf's law for cities: An empirical examination. *Regional Science and Urban Economics, 33*, 127–137.

Jin, B., & Wang, B. (1999). Chinese Science Citation Database: Its construction and application. *Scientometrics, 45*, 325–332.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*, 317–323.

Rousseau, R. (1992). Two remarks on the preceding paper by L. Egghe. *Information Processing and Management, 28*, 45–51.

Rousseau, R. (1994). Double exponential models for first-citation processes. *Scientometrics, 30*, 213–227.

Rousseau, R., & Rousseau, S. (1993). Informetric distributions: A tutorial review/Les distributions informétriques: un apercu. *Canadian Journal of Information and Library Science/Revue canadienne des sciences de l'information et de bibliothéconomie, 18*, 51–63.

Rowlands, I. (2002). Journal diffusion factors: A new approach to measuring research influence. *Aslib Proceedings, 54*, 77–84.
Schubert, A., Glänzel, W., & Braun, T. (1983). Relative citation rate: A new indicator for measuring the impact of publications. In D. Tomov & L. Dimitrova (Eds.), *Proceedings of the First National Conference with International Participation on Scientometrics and Linguistics of Scientific Text* (pp. 80–81). Bulgaria: Varna.
Thelwall, M. (2000). Web impact factors and search engine coverage. *Journal of Documentation, 56*, 185–189.
Wasserman, S., & Faust, K. (1994). *Social network analysis*. Cambridge, UK: Cambridge University Press.
Wilson, C. S. (1999). Informetrics. *Annual Review of Information Science and Technology, 34*, 107–247.
Wu, Y., Pan, Y., Zhang, Y., Ma, Z., Pang, J., Guo, H., et al. (2004). China Scientific and Technical papers and Citations (CSTCP): History, impact and outlook. *Scientometrics, 60*, 385–397.
Zipf, G. K. (1941). *National unity and disunity. The nation as a bio-social organism*. Bloomington, IN: Princeton Press.
Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.