Scientometrics, Vol. 30, No. 1 (1994) 35-47

BRIDGING THE GAPS: CONCEPTUAL DISCUSSIONS ON INFORMETRICS*

L. EGGHE

LUC, Universitaire Campus, B-3590 Diepenbeek (Belgium) +
and
UIA, Universiteitsplein 1, B-2610 Wilrijk (Belgium)

(Received January 18, 1994)

In this paper we discuss the possible gaps between several subdisciplines in informetrics and between informetrics and other -metrics disciplines such as econometrics, sociometrics and so on. It is argued that in all these disciplines, common models exist which describe the main points of interest. We also show that many concrete problems in these disciplines can be formulated in the same way and hence have similar solutions. We can conclude with the statement that the possible gaps between these disciplines are smaller than what many researchers in these different areas may feel and hence that many research projects could be set up in a wider framework.

Introduction

Before we can study "gaps" – as mentioned in the title, we must ask ourselves: "gaps *between what*?" Certainly between subdisciplines in the information sciences. How to define disciplines in information science depends on the point of view one has or one wants to highlight. One possible vision starts from a broad view on "information" in which one studies "potential information" (ordered or not) in relation with its "potential users". Indeed, only when potential information is really used one can talk about information: what would be the benefit of a library that is permanently closed or of a CD-ROM disc without a player?

The study of all kinds of statistical or mathematical aspects of information in this broad sense is called *informetrics*. Information in the above vision could be subdivided as in Fig. 1.

---

* Paper presented at the Fourth International Conference on Bibliometrics, Informetrics and Scientometrics in Berlin (Germany), September 11-15, 1993.

+ Permanent address.

Elsevier, Amsterdam-Oxford-New York-Tokyo
Akadémiai Kiadó, Budapest

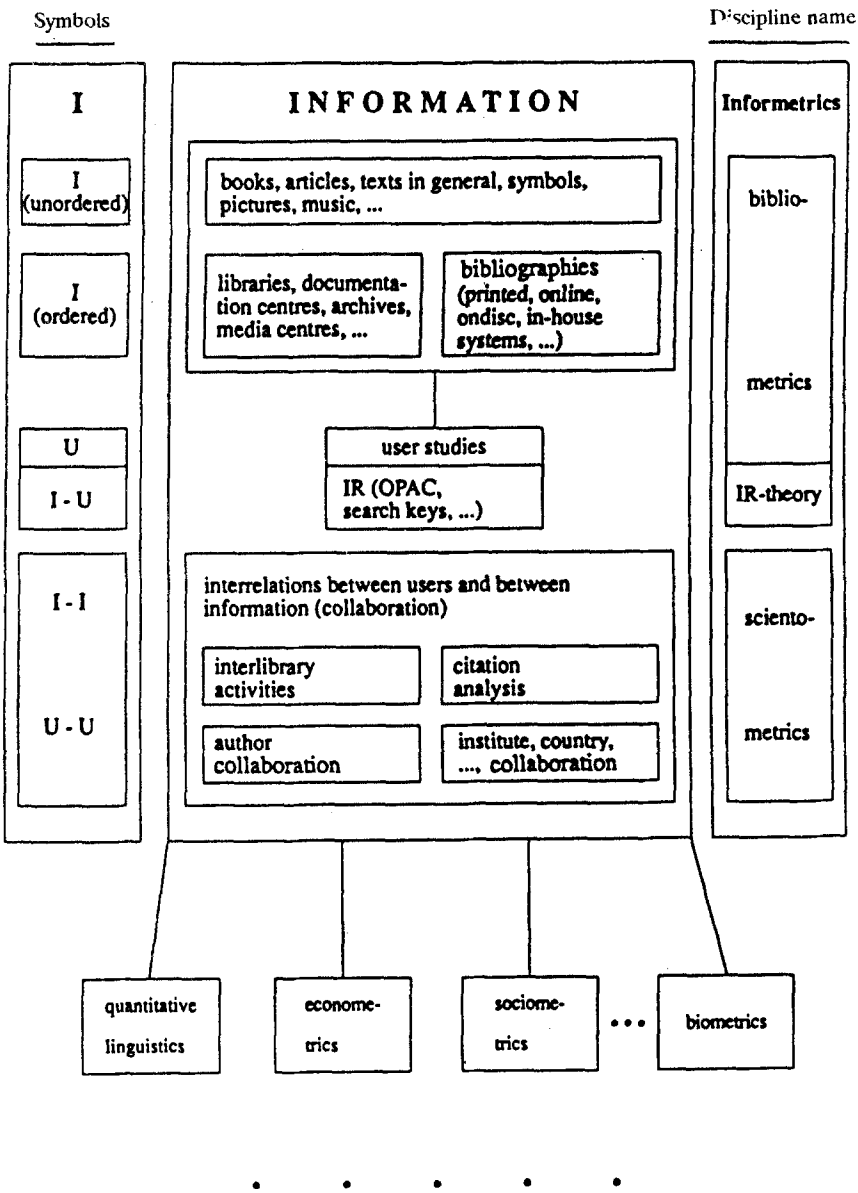


Fig. 1. Informetrics subdivision

One has unused – information studies (i.e. the documents as such, without users involved) which can be subdivided into unordered information and ordered information. Once users of information are involved one has user-studies and user-information studies (e.g. IR). These subdivisions of information lead to quantitative studies which could be called *bibliometric studies* (except – perhaps – for IR, i.e. information retrieval, they prefer their own terminology). Finally there are studies of the interrelations between information and between users such as citation analysis or authorcollaboration. This discipline is called *scientometrics*.

I do not want to go into any discussion on this division of "-metric" studies but I am well aware of the fact that other views (opinions) are possible. It is not important for the sequel. In fact the same can be said of the entire figure above: it represents one vision in which I can work out my ideas on "gaps" and the bridges to overcome them. I am only going to talk about the gaps as presented in the above figure. I am not going to discuss the gaps between

- theoretical studies and practical studies,
- the researchers themselves (librarians, professors, degrees in humanities, sciences, ...),
- sociological studies of information and mathematical ones,
- psychology and computer science (e.g. in information retrieval).

Occasionally I will also talk about the gaps between "Information science" (in the broad sense) and related major disciplines as quantitative linguistics, econometrics, sociometrics, biometrics and possible other "-metrics" disciplines.

The purpose of this paper is not to be exhaustive in the sense that all interrelations between the subdisciplines in "Information science" or between this discipline and other ones as mentioned above will be discussed. Our methodology will be of giving a number of very important key-words (as e.g. overlap, growth, obsolescence, ...) and then discuss these topics in the framework that was set up above.

We will not deal with techniques of data gathering and fitting (statistics) which are common to these disciplines nor with general mathematical methods (such as optimisation methods) which only serve as a technical helpware for these disciplines upon which mathematically explained models can be built. Hence our starting point is the "topic", not the "technique".

Finally we hope to have shown that many important topics can be defined exactly in the many subdisciplines of information science as well as in other disciplines, provided they are modified as necessary and consequently that studies on these topics can be executed in a broader way than they used to be. These considerations should be able to bridge the existing "gaps". This is the main purpose.

Common topical studies

IPP's (Information Production Processes)

In my Ph. D. (see *Egghe* (1989)) I developed the notion of "Information Production Process" (IPP), a term that was suggested by my supervisor, the late professor B. C. (Bertie) Brookes. Essentially an IPP is a "generalised bibliography" where one has sources (the producers) and items (the ones that are produced). The framework of IPP goes, however, far beyond the limits of information science, as the next table shows.

Sources	Items
journals	articles
authors	papers
articles	citations (giving or receiving)
books	their borrowings
search keys	documents with the same search keys
words (type)	their occurrence in a text (token)
employees	their salaries
cities	their inhabitants

In all these cases one can make a picture as in Fig. 2.

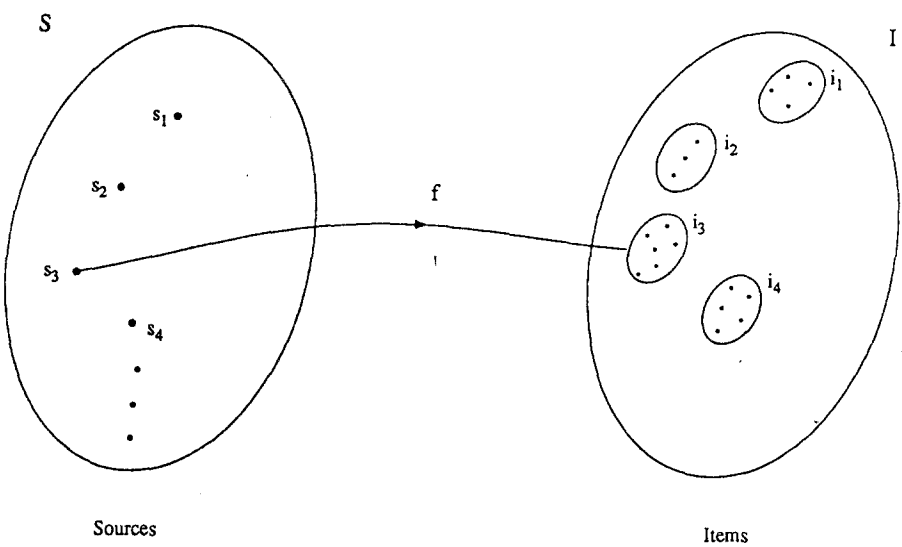


Fig. 2. An IPP

S = set of sources, I = set of items and

$$f:S \rightarrow 2^I$$

is the "device" function saying what sources have (or produce) which items. (2^I denotes the set of all subsets of I).

Central to the idea of IPP is the so-called duality between sources and items: there is a certain equivalence between sources and items: papers are written by authors and authors write papers. We remark here that duality is also a standard model in mathematics: it is encountered in geometry as well as in category theory [see (*Rousseau*, 1992b)]. Both mechanisms can be modeled and studied separately but the nice thing is that certain relations between one model and its dual exist: in a paper with 3 authors, each author could get a "weighted number" of produced papers of 1/3. This simple remark is on the basis of a dual theory relating author productivity models (number of papers per author) with collaborative models (number of authors per paper (see e.g. (*Egghe*, 1993b)).

It is remarkable that the author-paper relation in this sense is rather unique: here an item (paper) can have several sources (authors). This is not the case in the journal-article relationship: an article is published in *one* journal, a borrowed book is trivally uniquely related to *this* book, and so on. In the article-citations relationship, multiple sources for an item are also possible: a paper (source) can receive (give) many citations (items) and hence, when interchanging "receive" and "give", we have that an item can have many sources! It is our feeling that this special feature is not studied thoroughly so far.

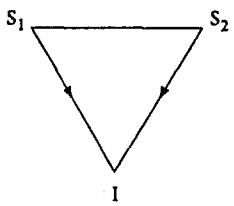
The so-called "success-breeds-success" (SBS)-principle can play an important role here (from partial results I have so far I know already that SBS can be generalized considerably in order to comprise more general IPP's (for references on SBS we refer the reader to *Ijiri* and *Simon* (1977) and *De Solla Price* (1976)).

Although borrowings are related to one specific book, the above observations could be applied here yielding new informetric studies, e.g. studying the borrowing behaviour of sets of books as a unity. Also one could study the co-occurrence of a set of several words in a text and so on.

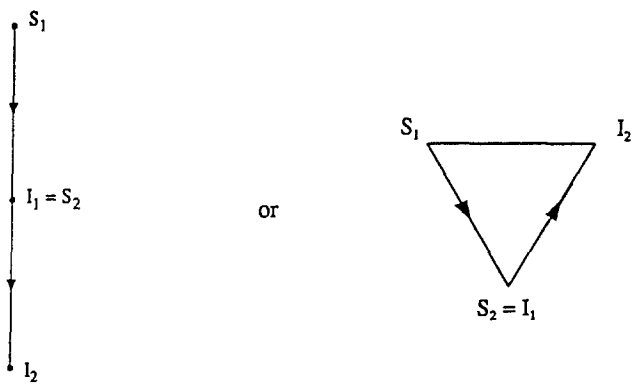
The links between (e.g.) the above mentioned disciplines are also evident from the fact that they often have the same (rank-) frequency laws (e.g. the laws of Bradford, Lotka, Zipf, Leimkuhler, Mandelbrot – some of them are mathematically equivalent). Indeed think of the mathematically, equivalent laws of Bradford

(*Bradford* (1934)) originally stated for bibliographies, Lotka (*Lotka* (1926)) stated for author productivities, Mandelbrot-Zipf (*Mandelbrot* (1977), *Zipf* (1949)) stated for texts and Pareto (*Pareto* (1895)) in econometrics.

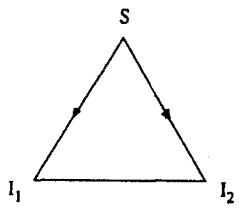
Another way to "bridge" the gaps is to study 3-dimensional IPP's. We will explain this now. Suppose we study the length of articles in a certain discipline (possibly as a function of time). This could be called a 1-dimensional study: only properties of one attribute (an article here) are involved. IPP's as explained above (source-item relationships) clearly involve two attributes and hence could be called 2-dimensional informetric studies. One can, however, remark that e.g. articles are written by authors and published in journals. Here we have two source sets and one item set. Such studies could be called 3-dimensional studies. The pictogram of the above described situation is (→ means "produce")



Other 3-dimensional studies can be done: journals have articles and articles have references. Here the 3-dimensional study looks as in the next figure:



Finally, one source can yield two item sets as e.g. in the case of articles that give references but also that receive references. The pictogram is as follows



Even 4 or higher-dimensional IPP studies are thinkable. It is clear that such studies (which hardly exist) give enormes possibilities for "gap bridging".

Finally we can also remark that IPP's studied in this way are related with fractal theory. In *Egghe* (1989), based on an argument of Mandelbrot (see *Mandelbrot* (1977)) we showed that two dimensional IPP's in the linguistical framework have fractal dimensions between 1 and 2, as it should! We conjecture that this will be true for any two dimensional IPP and that n-dimensional IPP's will have fractal dimensions between n-1 and n (notice the remarkable analogy with n-dimensional figures in Euclidean space).

Overlap

Overlap is a very important problem in many informetric topics. Yet the study is far from easy and even defining what overlap means requires careful thought.

In general, it seems easy to define overlap: it is the relative intersection of two sets (relative w.r.t. one of these two sets): Let A and B to two sets. Define the overlap of B w.r.t. A as

$$O(B|A) = \#(A \cap B)/\#A \qquad (1)$$

(i.e. the conditional probability to be in B, supposing we are in A: $P(A|B)$). This is a clear definition as long as we can determine $A \cap B$. Statistically this can be estimated by taking appropriate samples but this is not the main problem. The problem is to define "equal" members. If library A and B have the "same" book but in different editions, is this an overlap or not? Sometimes *yes* (e.g. were title catalogues are concerned), sometimes *no* (when the new edition is revised the user will consider the two books as different). How to define overlap in journal collections when the volumeholdings of the libraries are different? How to define overlap in research

projects? How to define overlap between two databases (e.g. MEDLINE and EMBASE)? Overlap is very interesting from a mathematical point of view: how to build union catalogues in order to optimise the work and time delay (we deal here with overlap problems between several collections!) (see e.g. *Buckland, Hindle* and *Walker* (1975))?

In all disciplines, a large overlap as well as a small overlap can be interesting (cf. also (*Egghe* and *Rousseau*, 1990)).

- Between libraries: the smaller the overlap the more new material is added (virtually) to a library when making an online connection with the other one. However, these new topics might not be of interest to this library. When the library is small; it might be interesting to look for a large library which covers the collection of the small library for about 100%.
- The same can be said about bibliographic databases.

About the construction of union catalogues one has the same two viewpoints:

- or one wants to cover as much library holdings as possible. In this case big overlap is more economic since it saves place (less titles),
- or one wants to cover as many titles as possible. In this case small overlap is more economic (less holdings).

Note again the dual viewpoint here.

Concentration, diversity

Concentration measures are functions that measure the degree of unequal division of the items over the sources. Usually such a function f is 1 in the perfectly concentrated case (one source has all the items) and 0 in the opposite case (all sources have an equal number of items). In addition to these requirements other "natural" properties are required for a function f to be a good concentration measure (e.g. the transfer principle, requiring that if one takes away from a poor person and give it to a rich one, concentration must increase – note the econometric terminology here!). Diversity measures are then functions g such that 1-g is a concentration measure. Hence no special theory for diversity measures is required; it can be derived immediately from the one of concentration measures. Well-known examples of good concentration measures are: the variation coefficient $V = \sigma/\mu$ and the Gini index G (G was introduced in 1909 in *Gini* (1909) in econometrics). In *Pratt* (1976), one had introduced a "new" concentration measure C, but *Carpenter* (1979) showed that C was nothing but a small variant of the Gini index:

$$C = [N/(N-1)]G \qquad (2)$$

where N is the number of classes. Since N is very high, we can say that $C \approx G$. This is a typical example of a "gap" between informetrics and econometrics that has been bridged in 1979 by *Carpenter*!

Concentration studies originate from econometrics since, there, a lot of "unequal" situations exist (e.g. the unequal division of wealth in the word and in each country and even in each community). It is well known that most situations in informetrics are unequal as well: few journals have many articles on a certain subject and many journals have few. The same with authors and publications. The same with words and their occurrence in texts (although here it are the least meaningful words that have the smallest ranks!). The same with cities (villages) and their inhabitants. The same with articles and their citations. The same with books and their use (e.g. borrowing) in a library.

Common results are therefore to be expected, the same good concentration measures can be used in all these disciplines: a clear bridge between the gaps!

Rousseau has done important work on concentration and diversity measures, thereby studying the literature of econometrics and ecology as well.

Yet it is still his opinion that many gaps between informetrics, ecology and econometrics exist. These gaps cry for bridges! For some gap bridging activities on this domain, see (*Rousseau*, 1992a) and (*Rousseau* and *Van Hecke*, 1993).

Ageing and growth

Ageing and growth are important topics in all informetric activities but they are also intimately related to each other. There are two reasons for that:

1. First of all the methodology for studying ageing (= obsolescence) and growth are the same: one studies (in the discrete version of time: years 1, 2, 3,...) the number of items in year t+1 divided by the number of items in year t, symbolically:

$$a(t) = : c(t+1)/c(t) \qquad (3)$$

For ageing, c(t) is the number of references that are t years old (synchronous study) or the number of citations, t years after publication (diachronous study).

Note the relation with scientometrics here. For growth, c(t) is the number of documents in year t (cumulative or not).

If in (3), a(t) is independent of t, a(t) = a we have exponential growth (a > 1) or exponential obsolescence (a < 1), but more sophisticated models exist (see e.g. *Egghe* and *Rao* (1992a, 1992b)). In *Egghe* (1994) it is shown (making relation with the Weber-Fechner sensation law) that, in the continuous setting, instead of eq (3) it is better to work with the formula

$$a(t) = e^{c'(t)/c(t)} \tag{4}$$

in all circumstances.

2. Secondly, there is a direct influence of growth on obsolescence: the larger the growth, the faster new literature is produced and hence the faster old literature has been used; however more and more literature becomes useable. Both aspects have an opposite influence on each other and it is not at all clear what will be the overall outcome. In *Egghe* (1993a) it is shown that, supposing exponential distributions for growth as well as obsolescence, that the larger the growth, the faster literature becomes obsolete (in the synchronous case) and that the opposite effect is true in the diachronous case.

Formula (3) (or (4)) is the basis for all growth or ageing studies and forms the underlying key for determining the ageing or growth distribution.

Much more than statistical fits of c(t) is the study of (3) or (4) revealing the true nature of the ageing or growth distributions (see *Egghe* and *Rao* (1992a, 1992b)). Note that many references on growth and ageing can be given (lots of them appear in the given references here; we do not emphasize a review paper here nor exhaustivity in the references).

Via (3) or (4) one can study the growth of libraries, the ageing of the library materials (e.g. by calculating the number of times a book is checked out from year to year) (i.e. local obsolescence), the growth (cumulative or not) of bibliographies, the growth of the scientific community (# researchers, ...). In general, one can study IPP's in function of time and see if decisions have to be changed in order to maintain optimal performance. A nice example of this are the so called search keys (introduced by Kilgour, the founder of OCLC – see *Kilgour* (1968)): a certain search key can be optimal (amongst the ones with the same length) but this optimality could be destroyed when the library catalogue continues growing. Of the same nature is the following: the larger a database,

the more time it takes to identify the requested items: are search techniques refining in the same pace in order to overcome this problem? Note that we are not talking here about the speed of computers: whether or not one uses a fast computer, the system will give the same bulk of information as reply to the same request (only the speed of deliverance of this bulk of information can be altered).

Furthermore it can be remarked that growth of the world literature has its influence on the growth of library collections and database sizes. The same can be said about ageing. In addition to that, ageing of literature is an important variable based on which we can determine library acquisitions: facing limited acquisition budgets one better focusses on literature that has larger ageing rates a (i.e. that becomes obsolete after a relatively long time period).

Codes

The inclusion of this topic in this overview has the purpose to show the link between informetrics and quantitative linguistics and to go back to the basic aspects of information theory. Textual information is trivially linked with linguistics and the quantitative study of it must have its influence in informetrics. The well-known law of Zipf (see *Zipf* (1949)), originating from linguistics has had many implications and uses in informetrics. One important application of Zipf-like regularities are in the area of compression theory. Knowing the exact regularity of word occurrence has influence on the code length one is going to use and hence an optimal compression can be reached.

Codes are important in IR, communication theory, libraries and, in fact, in virtual any aspect of information theory and even informatics (i.e. computer science). Codes, comprising the most elementary pieces of information (the bits) really form the basis of information theory. The Shannon-Weaver theory forms a model-theoretic basis for it (see *Shannon* and *Weaver* (1949)).

Codes, serving libraries and book production (e.g. ISBN) and IR (e.g. CODEN, search keys, ...) including check-digits are extremaly important. These check-digits are also used in ISBN, in CD-ROM technology and – outside our scope – as banking codes. Cryptography (secrets codes) are used in IR and communication theory (using advanced arithmetic methods). Compression is space – hence money-saving in IR, CD-ROM-technology and so on and is the main impetus for multimedia developments we are experiencing nowadays. In a way, coding theory forms the

elementary root from which the diverse disciplines in information science (and beyond) are developed.

Conclusions

I hope, with the overview, to have made it clear that there is an intimate methodological link between the different subdisciplines of informetrics and between informetrics and the other "-metrics" studies. By "methodological link" we mean that the same type of problems occur and that analogous ways of reasoning lead to solutions in all these (sub-) disciplines. Every problem in all these (sub-) disciplines needs, however, a dedicated definition of variables (cf. A. *Bookstein*, in his Bangalore paper (*Bookstein* (1992)). Once this problem is solved, the bridging of the gaps is possible. We note however that only "explained" arguments are able to provide such bridges. It is hence clear that purely statistical arguments cannot provide for this since it is always possible to fit different regularities by the same distributions. Statistics is only an auxiliary technique, preparing the explanations (let I make myself clear: statistics is very important in informetrics. It is *the* ultimate technique for librarians or more generally information brokers to evaluate their activities. Only in relation to the problem of "bridging the gaps" in this paper, statistics plays a ancillary role and has no explaining capabilities). With "explanations" I mean that, once the variables are defined accurrately, only logically deduced arguments (as e.g. in mathematics, probability theory and so on) can lead to gap-bridging results (modeling).

It is our sincere hope that more gap-bridging studies will emerge in the near future. It is then hoped that the authors of these papers are aware of the possible wider applicability of their arguments and hence that they adapt their papers in this way that they are general enough to be interpreted by researchers in the different (sub-) disciplines of informetrics and even beyond.

*

The author is grateful to Prof. Dr. R. *Rousseau* for interesting comments on this paper.

References

A. BOOKSTEIN (1992), Theoretical properties of the informetric distributions: some open questions. In: *Informetrics-91. Selected papers from the third International Conference on Informetrics*, Bangalore (India), I. K. R. RAO (Ed.), 1991, p. 17–35.

S. C. BRADFORD (1934), Sources of information on specific subjects. Engineering, 137, p. 85–86. Reprinted in: *Collection Management*, 1, p. 95–103 (1976–1977). Also reprinted in: *Journal of Information Science*, 10, 148 (facsimile of the first page) and 176–180 (1985).

M. K. BUCKLAND, A. HINDLE, G. P. M. WALKER (1975), Methodological problems in assessing the overlap between bibliographical files and library holdings. *Information Processing and Management*, 11, 89–105.

M. P. CARPENTER (1979), Similarity of Pratt's measure of class concentration to the Gini index. *Journal of the American Society for Information Science*, 30, 108–110.

D. DE SOLLA PRICE (1976), A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27, 292–306.

L. EGGHE (1989), *The duality of informetric systems with applications to the empirical laws*. Ph. D. thesis. The City University, London, UK.

L. EGGHE (1993a), On the influence of growth on obsolescence. *Scientometrics*, 27(2), 195–214.

L. EGGHE (1994), A theory of continuous rates and applications to the theory of growth and obsolescence rates. *Information processing and Management*, 30(2), 279–292.

L. EGGHE (1993b), Consequences of Lotka's law in the case of fractional counting of authorship and of first author counts. *Mathematical and Computer Modelling*, 18(9), 63–77.

L. EGGHE, I. K. RAVICHANDRA RAO (1992a), Citation age data and the obsolescence function: Fits and explanations. *Information Processing and Management*, 28 (2), 201–217.

L. EGGHE, I. K. RAVICHANDRA RAO (1992b), Classification of growth models based on growth rates and its applications. *Scientometrics*, 25 (1), 5–46.

L. EGGHE, R. ROUSSEAU (1990), *Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science*. Elsevier, Amsterdam.

C. GINI (1909), Il diverso accrescimento delle classi sociali e la concentrazione della richezza. *Giornale degli Economisti*, serie 11, 37.

Y. IJIRI, H. A. SIMON (1977), *Skew Distributions and the Sizes of Business Firms*. North-Holland, Amsterdam.

F. G. KILGOUR (1968), Retrieval of single entries from a computerized library catalog file. *Proceedings of the American Society for Information Science*, 5, 133–136.

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

B. MANDELBROT (1977), *The Fractal Geometry of Nature*. Freeman, New York.

V. PARETO (1895), La legge della domando. *Giornale degli Economisti*, ann. 12, 59–68.

A. D. PRATT (1977), A measure of class concentration in bibliometrics. *Journal of the American Society for Information Science*, 28, 285–292.

R. ROUSSEAU (1992a), *Concentration and diversity in informetrics research*. Doctorate thesis, University of Antwerp.

R. ROUSSEAU (1992b), Category theory and informetrics: information production processes. *Scientometrics*, 25 (1), 77–87.

R. ROUSSEAU, P. VAN HECKE (1993), Introduction of a species does not necessarily increase diversity. *Coenoses*, 8 (1), 39–40.

C. E. SHANNON, W. WEAVER (1949), *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois, USA.

G. K. ZIPF (1949), *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge. Reprinted by Hafner, New York (1965).