EXPANSION OF THE FIELD OF INFORMETRICS: ORIGINS AND CONSEQUENCES

by

L. Egghe

Limburgs Universitair Centrum (LUC), Universitaire Campus, B-3590 Diepenbeek, Belgium (*)

and

Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium

(*) Permanent address

Key words and phrases: informetrics, scientometrics, bibliometrics, webometrics, growth

---

ABSTRACT

This editorial introductory paper first discusses the reasons for the clear growth of the field of informetrics (bibliometrics, scientometrics, webometrics, ...). This has lead some journals to increase their number of volumes or the number of issues per volume. The journal Information Processing and Management decided to devote two special issues (the one here and another one to come in 2006) to the broad topic "Informetrics" where the scope of these special issues is to attract good papers dealing with gathering important data sets and/or presenting original models and explanations. Then we briefly discuss the content of the papers that are published in this special issue. They are dealing with models, mapping of science (cocitation, coword analysis), web sites and search engines, collaboration in digital libraries and the newest topic in informetrics: use of and access to articles in digital libraries.

I. THE GROWTH OF THE FIELD OF INFORMETRICS

In this introductory paper, we will use the term "informetrics" as the broad term comprising all -metrics studies related to information science, including bibliometrics (bibliographies, libraries, ...), scientometrics (science policy, citation analysis,

research evaluation, ...), webometrics (metrics of the web, the Internet or other social networks such as citation or collaboration networks), ...

The term informetrics was introduced by Blackert and Siegel (1979) and by Nacke (1979) but gained popularity e.g. by the organization of the international informetrics conferences in 1987 (see Egghe and Rousseau (1988, 1990)). However the field "informetrics" (not the name) started already in the first half of the twentieth century e.g. by the works of Lotka, Bradford and Zipf (see Lotka (1926), Bradford (1934), Zipf 1949, but for the law of Zipf, see also Condon (1928) or even Estoup (1916)). The term bibliometrics was coined in Pritchard (1969) and the term scientometrics was coined in Nalimov and Mul'čenko (1969) in Russian: naukometrija. For more on the history of these and other terms see White and McCain (1989), Ikpaahindi (1985), Lawani (1981), Tague-Sutcliffe (1994), Brookes (1990), Wilson (1999), Egghe and Rousseau (1990) and Egghe (2005).

That the field of informetrics has grown in the twentieth century is evident but this growth has become more and more clear the last decades. Lipetz (1999) describes an exponential growth of JASIS - now called JASIST (Journal of the American Society for Information Science and Technology, existing 50 years in 1999) in terms of number of papers and in terms of number of authors and even in terms of average number of references per paper. One also shows in Lipetz (1999) that the average number of authors per paper is increasing.

Authors are also responsable for a multidisciplinary growth of the field of informetrics - see Summers, Oppenheim, Meadows, McKnight and Kinnell (1999) hereby also indicating the influence of informetrics to other scientific disciplines. Multidisciplinarity is evident if one looks at the "new" topics which informetrics is covering: the metrics of the web, Internet, intranets and other social networks such as citation or collaboration networks. In general one can say that the creation of the "information society" is responsable for the growth of the field of informetrics. So we can say that the field of informetrics nowadays comprises the fastly growing field of webometrics (see Hood and Wilson (2001)) (netometrics, as introduced in Bossy (1995) would be a better term covering also non-web activities but the term does not seem to become popular - see Hood and Wilson (2001)). Cybermetrics also exists (it is even the name of an electronic journal under the editorial direction of I. Aguillo) but it is not clear whether it will overtake, some day, the term webometrics.

Schubert (2002) describes 50 volumes of the journal Scientometrics and also concludes the increase of the number of authors and the fact that they more and more collaborate in the sense that the average number of authors per paper increases (same conclusions as in Lipetz (1999)). Schubert also remarks that there is no evidence that the degree of "hardness" of the field informetrics is increasing, a point to keep in mind for the future evolution of this field. He and Spink (2002) describe foreign authorship in JASIST and JDOC (Journal of Documentation) and prove that their share in these journals becomes larger and larger indicating an increase of internationalization of the field of informetrics. The latter is also illustrated in Bar-Ilan (2000) where one makes the constatation that the articles in the Proceedings of the international informetrics conferences are increasingly cited.

The extension of information science to networks and the information society in general has the consequence that more and more data are gathered in an automatic way. This implies that data can be gathered in a much faster way than it used to be but also that the accuracy is dropping. There are several reasons for this. First of all one gets data from a documentary system (e.g. an OPAC, secondary or primary

electronic database or digital library) but, since there is - in general - no clear definition of the topics due to lack of standards (see Glänzel (1996), Rousseau (2002)) one is not completely sure of what one gets. In addition an electronic system may suffer from system breakdown in which case one is obliged to make unexact interpolations.

Data of electronic services and activities through the web (many data are) are also of a different nature than data gathered directly from a computer system. An example is connect time versus times of connection. When entering directly or via telephone lines into a computer system (e.g. an OPAC or the DIALOG system) one is able to report on the connect time. When using a documentary system via the web one cannot report on connect time anymore but only on number of connections (cf. the well-known DIALOG units). Networks such as the web typically have connections between the sites and one talks in this connection about hyperlinks (in-links when a site receives a hyperlink from another site; out-links when a site gives a hyperlink to another site). Their informetric distributions have been studied even in journals such as Nature and Science (see e.g. Albert, Jeong and Barabási (1999), Barabási and Albert (1999) and Huberman, Pirolli, Pitkow and Lukose (1998)) but also in physics journals (see e.g. Barabási, Jeong, Néda, Ravasz, Schubert and Vicsek (2002) and Adamic, Lukose, Puniyani and Huberman (2001)), again showing the interdisciplinary character of nowadays informetrics. Hyperlinks usually are compared with the better known citations but they are very different of nature: hyperlinks cannot be used for aging or author collaboration studies since they are not dated and are - usually - anonymous. Hyperlinks can be used for determining "authoritative" web sites or documents - see CLEVER (1999) which in turn can be used in information retrieval (IR). Also in IR, quantitative methods, e.g. for the evaluation of searches and systems have drastically changed by the way search engines deliver search results: they give the retrieved documents in decreasing order of expected relevance which creates the need for evaluation measures on ordered sets instead of he classical ones (e.g. recall, precision, Jaccard, Cosine, Dice, ...) on ordinary sets (cf. Egghe and Michel (2002, 2003)).

It is very important to mention that the fact that most articles are nowadays appearing in electronic journals and/or repositories gives the new possibilities of measuring the use of articles not only by citations or web citations but also by measuring their number of downloads. Downloads can be considered as electronic versions of reading or photocopying of a paper article. The latter indicators were never studied due to the great difficulty of manual datagathering. Hence the study of downloads and their relation with (web) citations is intriguing, see Antelman (2004), Brody and Harnad (2004), Harnad and Brody (2004a,b) and Perneger (2004).

It is clear from the above that the extension of informetrics to electronic - e.g. web - activities gives a boost to the challenge of datagathering and datamanagement and hence to the growth of the field. The need for more publication outlet, which is a consequence from this, is also clearly seen if one looks at the two important informetrics journals JASIST and Scientometrics. JASIST decided in 1998 to increase its publication flow from 12 issues to 14 issues a year. Scientometrics is publishing, from 2005 onwards, 12 issues instead of 9 issues per year. In this connection I want to give a personal advise, which is shared with the informetric colleagues I contacted recently. The increase of publication outlet does also increase the need of refereeing. It is my personal feeling that one should expand the list of possible referees in informetrics to younger informetricians: my workload on refereeing has doubled in 2004, a phenomenon that is recognized by colleague informetricians.

Apart from JASIST and Scientometrics, the present journal Information Processing and Management (IPM) is the only journal that regularly publishes papers devoted to informetrics studies, although, in general, IPM is more focused to the subfield of informetrics dealing with quantitative aspects of IR. Elsevier, the publisher of IPM, is interested if a more pronounced general informetrics component is possible in IPM. Hereby one wants to stress that the principal goal is to give an outlet to high quality papers in informetrics. High quality papers are papers that present good mathematical (probabilistic) models and explanations of informetric regularities (in the broad sense) and/or papers in which interesting and important datagathering is presented. The former request (good models and explanations) can be understood in the framework of increasing the degree of "hardness" of the "science" informetrics (cf. Schubert (2002), as mentioned above, there is no evidence that the "hardness degree" has increased recently). The latter request (important datagathering) can be understood in the connection described above: the need for new informetric data coming from electronic environments such as the Internet, so that the regularities in these new media can be understood. Of course, important new data coming from "classical" informetric topics (e.g. cocitations) are also intersting.

The papers in this special issue were selected based on these two broad principles. In the next section we will present a brief description of these papers.

II. THE PAPERS IN THIS SPECIAL ISSUE

Models can be found in five papers. The paper of Burrell, entitled "Symmetry and other transformation features of Lorenz/Leimkuhler representations of informetric data", deals with econometric aspects of informetrics by studying the Lorenz curve. He proves that the Lorenz curve determines the production distribution and examins powers of Lorenz curves. Also self-symmetry aspects of Lorenz curves are studied. Also the paper of Egghe (independently refereed), entitled "Continuous, weighted Lorenz theory and applications to the study of fractional relative impact factors", deals with Lorenz curves. Here, relative impact factors, interpreted in the fractional way, are characterized by the construction of weighted Lorenz curves. Within this model Egghe shows that if, for two situations, one fractional impact factor is larger than the other one, the same is true for all other fractional impact factors and that this result is not true for "classical" impact factors using fixed time periods.

The paper of Rousseau, entitled "Conglomerates as a general framework for informetric research", generalizes the well-known "information production processes" (IPPs) by adding the notion of a pool and a magnitude map for item-sets. In this generalization, conglomerates apply to (web) impact factors, Bradford-Lotka type bibliographies, word use, diffusion factors, elections and even bestsellers lists. Generalized Zipf type distributions are studied in the paper of Shan, entitled "On the generalized Zipf's distribution. Part I". General Zipf type distributions are functions that show an approximately linear right tail on a log-log scale. Their characteristics are studied and these are used to describe Zipfian phenomena. The fifth paper on models is the paper of Lafouge and Prime Claverie, entitled " Links between entropy and production of information. Characterization of bibliometric distributions using the effort function". Here production distributions (such as the geometric (exponential) and the power model, i.e. Lotka's function) are characterized by corresponding effort functions and, in each case, the relation with the entropy is given.

Four papers deal with the "new" topic of use of and access to electronic articles in a digital library. In the paper of Kurtz, Eichhorn, Accomazzi, Grant, Demleitner,

Henneken and Murray, entitled "The effect of use and access on citations", the authors study the possible influence of use and access of articles prior to publication on later citations from the viewpoints: OA (Open Access), EA (Early Access) and SB (Self-selection Bias). Zhao's paper, entitled "Challenges of scholarly publications on the web to the evaluation of science - a comparison of author visibility on the web and in print journals", reveals different patterns of scholarly communication on the web and in print journals and promotes the idea of a "two tier" communication and evaluation system, complementing the Web of Science databases. A similar topic is addressed in the paper of Bollen, Van de Sompel, Smith and Luce, entitled "Toward alternative metrics of journal impact. A comparison of usage and citation data". They determine alternative journal impacts based on network centrality measures and conclude that the "classical" impact factors cannot be the sole assessment of journal impact, hence needing again a "two tier" system where also journal impact measures are used, based on usage data. The fourth paper in this subfield is of Nicholas, Huntington, Dobrowolski, Rowlands, Hamid Jamali and Polydoratou and is entitled "Revisiting "obsolescence" and journal article "decay" through usage data: an analysis of digital journal use by year of publication". Hence, as in the two previous papers, usage of articles in a digital library is taken as an alternative of citations but now to determine obsolescence or aging.

Collaboration (co-authorhip) is a classical subfield of informetrics. In this special issue we have two papers dealing with this topic but in the environment of web networks or digital libraries. The paper of Liu, Bollen, Nelson and Van de Sompel, entitled "Co-authorship networks in the digital library research community", deals with social network analysis applied on the co-authorship network of past digital library conferences. A variant of PageRank, AuthorRank is introduced and results are compared with other ranking techniques such as ranks based on network centrality measures. Indicators of gender centrality and bibliometric and web indicators of gender cooperation has been executed on the set of multi-authored publications of 64 COLLNET members in the paper of Kretschmer and Aguillo, entitled "New indicators for gender studies in web networks".

Further web studies are found in the following two papers. First there is the paper of Payne and Thelwall, entitled "Mathematical models for academic webs: linear relationship or non-linear power law ?". Here one shows, experimentally, that the relation between research of a university and links to the university's web site is a linear one and not a non-linear power law. Bar-Ilan, in the paper entitled "Comparing rankings of search results on the web", rank results of several IR commands are compared in Google, AlltheWeb, Alta Vista and HotBot and one concludes that the employed ranking algorithms are considerably different.

There are three papers dealing with the mapping of science. Moya-Anegon,Vargas-Quesada, Chinchilla-Rodríguez, Corera-Álvarez, Herrero-Solana and Munoz-Fernández have a paper entitled "Domain analysis and information retrieval through the construction of heliocentric maps based on ISI-JCR category cocitation". Based on the JCR Subject Categories and cocitation between them, they construct heliocentric maps of major scientific domains in Spain, France and England and the results are compared. Cocitation is also used in the paper of Marshakova-Shaikevich, entitled "Bibliometrics maps of field of science". Using again data from the ISI (Thomson) citation indexes, maps are constructed based on journal cocitation and lexical analysis of keywords in the titles and texts. The same source is used in the paper of Glenisson, Glänzel, Janssens and De Moor, entitled "Combining full-text and bibliometric information in mapping scientific disciplines". This combined methodology of text mining (coword analysis) and bibliometric techniques (cluster analysis) is applied to the papers in the 2003 volume of the journal Scientometrics.

Coword analysis is also applied in the last paper in this Special Issue. It is the paper of Onyancha and Ocholla, entitled " An informetric investigation of the relatedness of opportunistic infections to HIV/AIDS". Through the analysis of published articles one can show the disease-gene relationship, i.e. the relatedness of the AIDS-defining diseases in persons with documented HIV infection. Coword analysis is used to calculate the strength of association between the desriptors of the diseases and the gene.

REFERENCES

Adamic, L.A., Lukose, R.M., Puniyani, A.R. and Huberman, B.A. (2001). Search in power-law networks. Physical Review E, 64, 46135-46143.

Albert, R., Jeong, H. and Barabási, A.-L. (1999). Diameter of the World-Wide Web. Nature, 401, 130-131.

Antelman, K. (2004). Do open-access articles have a greater research impact ? College and Research Libraries, 65(5), 372-382.

Bar-Ilan, J. (2000). The web as an information source on informetrics ? A content analysis. Journal of the American Society for Information Science and Technology, 51(5), 432-443.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. Science, 286, 509-512.

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. Physica A, 311, 590-614.

Blackert, L. and Siegel, S. (1979). Ist in der wissenschaftlich-technischen Information Platz für die Informetrie ? Wissenschaftliches Zeitschrift TH Ilmenau, 25(6), 187-199.

Bossy, M.J. (1995).The last of the litter: "Netometrics". Solaris Information Communication, 2, 245-250.
http://www.info.unicaen.fr/bnum/jelec/Solaris/d02/2bossy.html

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

Brody, T. and Harnad, S. (2004). Earlier web usage statistics as predictors to later citation impact. http://www.ecs.soton.ac.uk/~harnad/Temp/timcorr.doc

Brookes, B.C. (1990). Biblio-, sciento-, infor- metrics ??? What are we talking about ? In: Informetrics 89/90. Proceedings of the second international Conference on Bibliometrics, Scientometrics and Informetrics (L. Egghe and R. Rousseau, eds.), 31-34, Elsevier, Amsterdam, the Netherlands.

CLEVER (1999). Hypersearching the web. Scientific American, June 1999, 44-52.
http://www.sciam.com:80/1999/0699issue/0699raghavan.html

Condon, E.U. (1928). Statistics of vocabulary. Science, 67(1733), 300.

Egghe, L. (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier, Oxford, UK.

Egghe, L. and Michel, C. (2002). Strong similarity measures for ordered sets of documents in information retrieval. Information Processing and Management, 38(6), 823-848.

Egghe, L. and Michel, C. (2003). Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques. Information Processing and Management 39(5), 771-807, 2003.

Egghe, L. and Rousseau, R. (eds.) (1988). Informetrics 87/88. Proceedings of the First International Conference on Bibliometrics and Theoretical Aspects of Information Retrieval. Elsevier, Amsterdam, the Netherlands.

Egghe, L. and Rousseau, R. (1990a). Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science. Elsevier, Amsterdam, the Netherlands.

Egghe, L. and Rousseau, R. (eds.) (1990b). Informetrics 89/90. Proceedings of the Second International Conference on Bibliometrics, Scientometrics and Informetrics. Elsevier, Amsterdam, the Netherlands.

Estoup, J.B. (1916). Gammes Sténographiques. 4th Edition. Institut Sténographique, Paris.

Glänzel, W. (1996). The need for standards in bibliometrics research and technology. Scientometrics, 35(2), 167-176.

Harnad, S. and Brody, T. (2004a). Comparing the impact of open access (OA) vs. non-OA articles in the same journals. D-Lib Magazine, 10(6). http://www.dlib.org/dlib/june04/harnad/06harnad.html

Harnad, S. and Brody, T. (2004b). Prior evidence that downloads predict citations. British Medical Journal Rapid Responses, September 6. http://bmj.bmjjournals.com/cgi/eletters/329/7465/546#73000

He, S. and Spink, A. (2002). A comparison of foreign authorship distribution in JASIST and the Journal of Documentation. Journal of the American Society for Information Science and Technology, 53(11), 953-959.

Hood, W.W. and Wilson, C.S. (2001). The literature of bibliometrics, scientometrics and informetrics. Scientometrics, 52(2), 291-314.

Huberman, B.A., Pirolli, P.L.T., Pitkow, J.E. and Lukose, R.M. (1998). Strong regularities in World Wide Web surfing. Science, 280, 95-97.

Ikpaahindi, L. (1985). An overview of bibliometrics: its measurements, laws and their applications. Libri, 35(2), 163-177.

Lawani, S.M. (1981). Bibliometrics: its theoretical foundations, methods and applications. Libri, 31(4), 294-315.

Lipetz, B.-A. (1999). Aspects of JASIS authorship through five decades. Journal of the American Society for Information Science, 50(11), 994-1003.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences, 16(12), 317-324.

Nacke, O. (1979). Informetrie: eine neuer Name für eine neue Disziplin. Nachrichten für Documentation, 30(6), 219-226.

Nalimov, V.V. and Mul'čenko, Z.M. (1969). Naukometrija. Nauka, Moskva, USSR.

Perneger, T.V. (2004). Relation between online "hit counts" and subsequent citations: prospective study of research papers in the British Medical Journal. British Medical Journal, 329, 546-547.
http://bmj.bmjjournals.com/cgi/content/full/329/7465/546

Pritchard, A. (1969). Statistical bibliography or bibliometrics ? Journal of Documentation, 25, 348-349.

Rousseau, R. (2002). Lack of standardisation in informetrics research. Comments on "Power laws of research output. Evidence for journals of economics" by Matthias Sutter and Martin G. Kocher. Scientometrics, 55(2), 317-327.

Schubert, A. (2002). The web of Scientometrics. A statistical overview of the first 50 volumes of the journal. Scientometrics, 53(1), 3-20.

Summers, R., Oppenheim, C., Meadows, J., McKnight, C. and Kinnell, M. (1999). Information science in 2010: a Loughborough university view. Journal of the American Society for Information Science, 50(12), 1153-1162.

Tague-Sutcliffe, J. (1994). Quantitative methods in documentation. In: Fifty Years of Information Progress: a Journal of Documentation Review, 147-188, Aslib, London, UK.

White, H.D. and McCain, K.W. (1989). Bibliometrics. Annual Review of Information Science and Technology (ARIST), 24 (M.E. Williams, ed.), 119-186.

Wilson, C.S. (1999). Informetrics. Annual Review of Information Science and Technology (ARIST), 34 (M.E. Williams, ed.), 107-247.

Zipf, G.K. (1949). Human Behavior and the Principle of least Effort. Addison-Wesley, Cambridge, USA. Reprinted: Hafner, New York, USA, 1965.