DISCUSSIONS ON INFORMETRICS OF THE INTERNET AND OTHER SOCIAL NETWORKS

by

L. EGGHE

LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium
e-mail : leo.egghe@luc.ac.be

Summary

This paper poses more problems than it solves : it investigates the new (virtual) world of Internet and the challenges that it offers for informetric analysis. The paper studies four different aspects. First of all there is the increasing problem of data gathering in the Internet. Second topic is the Internet-version of the informetric laws : are the same types of classical distributions valid or not ? Third topic deals with scientometric aspects : can the clickable buttons (hyperlinks) in Web pages replace the role of classical references in scientific papers ? It also contains a study of the WIF (Web Impact Factor) and a discussion on aging. The fourth topic discusses IR (information retrieval) aspects of search engines. It studies aspects of probabilistic IR as applied in these engines and poses the question of quantitative evaluation of IR (Web analogues of recall and precision or of Jaccard's measure, Dice's measure or cosine measure as applied to ordered sets of documents). Concrete

---

[1] Permanent address

formulas are presented of ordered similarity measures that are capable of comparing ordered sets of documents.

I. Introduction : data gathering.

I, as a conscious librarian for 20 years and as a mathematician for 25 years, have always been concerned with reporting on library actions and holdings. The needs for such reports (e.g. annual reports) are very clear :

- **they are needed to convince subsidising bodies (e.g. Rector, Minister, ...) for giving enough money and staff to the library**
- **they are needed to inform (p.r. and p.a. : public relations and public awareness) the library users on why things are organised the way they are and on why certain services are not free.**
- **they are needed for the library manager (chief librarian) as a managerial tool and as a source of (otherwise hidden) information.**

Indeed, it is typical for library actions - as is well-known by librarians - that many activities are "hidden" or at least not well known by external persons. A typical example is the heavy daily task of reshelving used books (in my library about 20,000 per year). Informing on such activities convinces external people of the high amount of work that has to be performed in a library.

In the last years, however, the number of "electronic" activities has increased drastically. In most cases this also means that data are gathered in an automatic way and hence one is inclined to think that it has become easier to collect data. This is not true. It is true that more and more data are gathered in a much faster way but at the same time their accuracy has dropped. One reason can be the fact that these data are delivered by the computer via a third person who might have another insight on what the exact definition of a certain attribute is. I have some experience in this matter : the LUC-library forms a network with the University of Antwerp libraries and the automation team is in Antwerp. So it happens that quantitative topics wanted by LUC and Antwerp are not

always exactly the same (and sometimes one even does not know the difference !). An example is the information of users (based on users' barcodes) : are all users counted or only the ones that were activated this academic year (i.e. the ones that used the library at least once this academic year). Another problem is to report on the number of books added to the collection this year : are free books included (e.g. theses), are new editions included, are multiple copies included, how are serials counted, and so on. The problems arise because the data are generated by a computer (and not by each librarian manually) and hence it is not easy to make sure that what is in the librarian's mind is also delivered in the same way.

Another reason for the increase of problems of data gathering in an automatic way is that, during the year one has some periods of system break down and hence one has loss of data. Sometimes it is not seen, sometimes it is seen and one applies a method of "interpolation" but in any case the final result is not exact. An example is given by not-registered circulations of books.

The problem of data gathering in an electronic environment got worse even more since more and more activities in libraries are web-oriented. A typical example is a web-OPAC (OPAC = Online Public Access Catalogue). My library catalogue has been automated in 1989 and has become a web-catalogue around 1995. Before this I was able to report on the search time in the library's OPAC. This is not possible anymore for the web-OPAC. A similar problem is experienced by DIALOG users. DIALOG is reachable via WWW. Scientists or librarians who use this link find out that there is no connect time indicated anymore (nor is it invoiced this way) ; one counts now with DIALOG units but there is no clear definition for it and even if there is one (I assume DIALOG people have a definition !) it cannot be used to measure connect time in a file.

We have come across the first major difference between the Internet (the virtual world) and the real world : in the latter "use" is measured by time ; in the former "use" is measured by number of times there has been

contact. It is not clear what the impact of such a big change will be on the (social) habits of information exchange.

Even "number of contacts" is sometimes difficult to measure or is a fuzzy notion. Let us go back to the example of the OPAC. Since it has become a web-OPAC, contact is possible from outside the library and even from any place in the world. It is therefore

- **not easy to report on the number of OPAC contacts**
- **not very relevant to report on all these contacts since an OPAC search from e.g. China to the LUC catalogue has a different goal than an OPAC search within the LUC library.**

We close this section by making the obvious remark that also the incredible size of the Internet (and its fast growth - see further) are an obstacle to perform searches and samples, needed in data gathering.

I think these few examples show the degree of complexity on reporting in a quantitative way on web (Internet) based activities. In the sequel we will address more "fundamental" problems in the sense that we will study new informetric aspects of this new information space.

II. Networks and (classical) informetric laws

II.1 Sources and items

One of the most evident questions that can be asked in this context is : Are the classical informetric laws valid in networks, e.g. the Internet ? In other words, are the webometric laws the same as the informetric ones ? This question was also posed by Boudourides, Sigrist and Alevizos (1999) but not at all answered by them !

Before one can answer this question one must look at the ingredients of the classical informetric laws in the real information world. Classical informetrics deals with sources (the objects that produce), items (the

objects that are produced) and a linking function f determining which items are produced by which sources. This framework was studied by Egghe (1989, 1990) in the connection of duality (between sources and items). The system of sources, items and linking function was called an IPP (Information Production Process). Classical examples are : bibliographies (sources = journals or authors, items = articles), citation lists (sources = articles, items = citations or references), texts (sources = word types, items = word tokens) and so on.

In this general setting it is easy to formulate the classical informetric laws such as the ones of Lotka, Bradford, Mandelbrot, Zipf, Leimkuhler, ... These laws are well-known but I will repeat them here (some of them).

Let us just repeat two basic laws: the ones of Lotka and Zipf.

The law of Lotka

"The number of sources f(n) with n=1,2,3,... items equals

$$f(n) = \frac{C}{n^{\alpha}} \qquad (1)$$

, where C and are constants, C, >0." Usually $1 and its "classical" value is =2 which is a "turning" point in informetrics.

The law of Zipf

"If we rank the sources in decreasing order of the number of items in these sources, then the number of items in the source on rank r=1,2,3,... equals

$$g(r) = \frac{D}{r^{\beta}} \qquad (2)$$

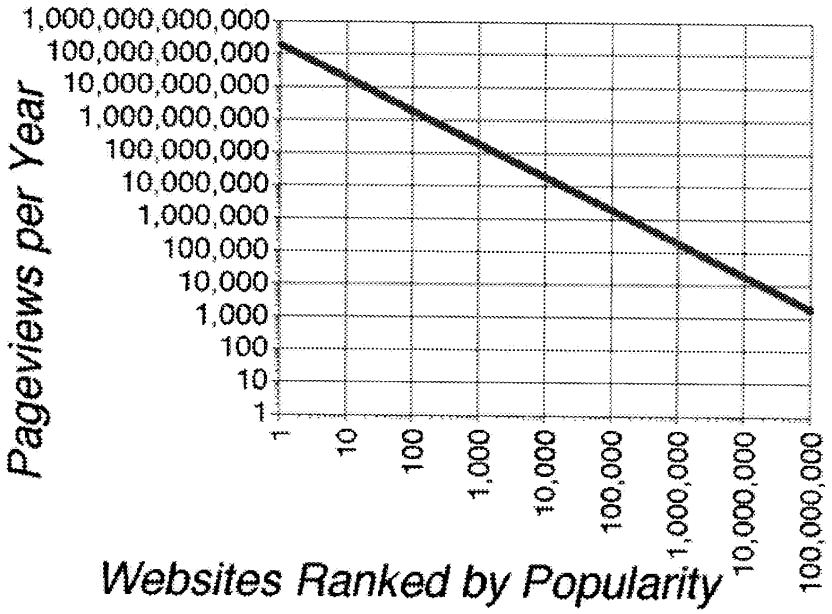
, where D and are constants, D, >0". It can be shown that Zipf's law implies Lotka's and the relation between their exponents is:

$$\beta = \frac{1}{\alpha - 1} \qquad (3)$$

For more information see Egghe (1989, 1990) or the book Egghe and Rousseau (1990).

II.2 Examples

So in order to be able to answer the question : "Are the classical informetric laws valid in the Internet ?" , a necessary requirement is that we are able to determine (each time) the sources and the items that are produced by these sources . This is, however, not always evident. Web pages do not always have an explicite author and are not published in a journal (except the ones published in an electronic journal). In the previous section we mentioned already that connect times are impossible to determine and that use (of e.g. web pages) is expressed in terms of number of logins (these play the role of items here). This feature has e.g. been studied in Nielsen (1997) where one produces an acceptable prediction of the cumulative number (= rank) of web sites (= sources) versus the number of pageviews (= items) per year for each site and this for the year 2000 (an update does not exist). This rank-frequency distribution clearly is Zipfian (see Fig. 1).



Predicted usage numbers for websites in the Year 2000.
The x-axis shows sites ranked by popularity (#1 is the most heavily used site)
The y-axis shows the number of pageviews per year for each site
Note that both axes have logarithmic scales

Fig. 1

The same conclusions are found in Adamic and Huberman (2002) on the ranking of websites according to visitors on December 1, 1997 via the provider America Online.

Other clearly defined source-item relations are : web sites and their size (# of pages), web pages (or sites) and their number of clickable buttons. The latter one is very interesting and will be revisited in the next section. There clickable buttons (also called hyperlinks) are compared with classical references in papers. We will, however, show that in this comparison also differences are present.

In Rousseau (1997) it is shown (in a statistical way) that the distribution of hyperlinks between web sites is of Lotka type. The value of the Lotka exponent is around 2.3. This also goes for the distribution of domain names

(such as .edu, .com, .uk, .cn, .fr, .be and so on). The Lotka exponent is around 1.5.

In Adamic and Huberman (2001,2002) (see also Huberman (2001)) one finds the following (see Fig. 2) power laws (Zipf) for the rank-frequency functions of rank of sites versus number of pages, number of users, number of out-links, number of in-links.

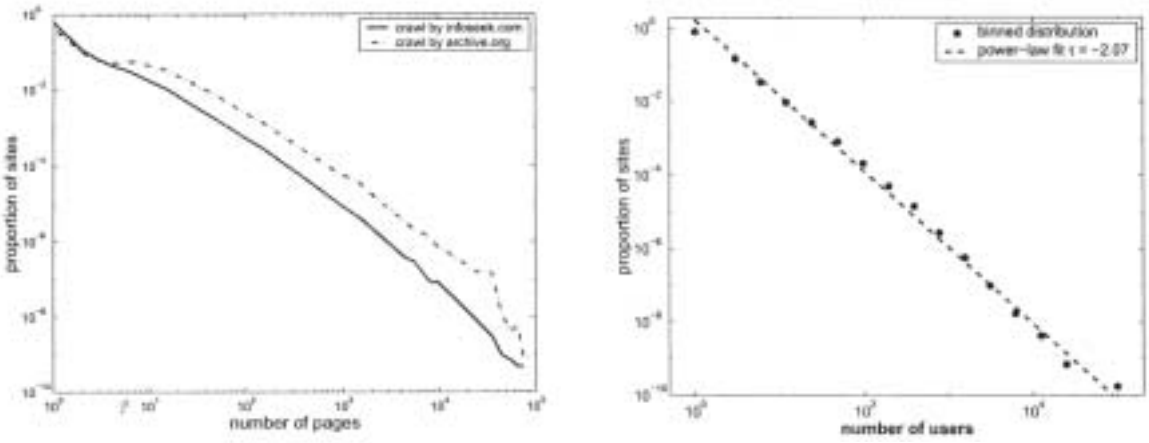
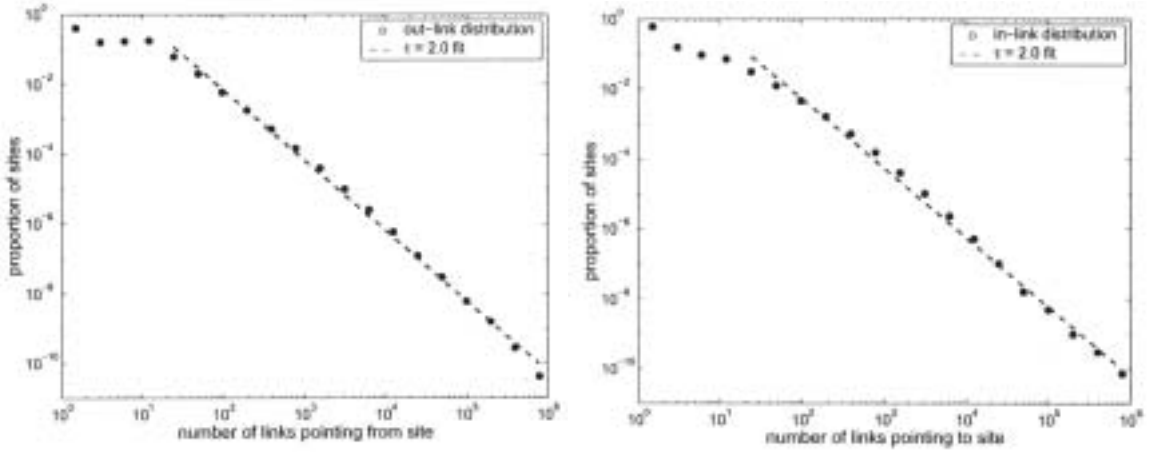


Fig. 2 Rank-frequency distributions for websites



In the following papers one considers general "social networks", comprising citation networks, collaboration networks, internet and WWW, intranets,...: Bilke and Peterson (2001), Jeong, Tombor, Albert, Ottval and Barabási (2000), Barabási, Jeong, Néda, Ravasz, Schubert and Vicsek (2002), Adamic, Lukose, Puniyani and Huberman (2001), Barabási and Albert (1999).

There one finds that social networks follow Lotka's law (for the number P(p) of points with p links) with high exponents between 2 and 4. Social networks are in contrast with random networks in which links between points are given in a random way. These so-called Erdös-Rényi networks (see Erdös and Rényi (1960)) do not even follow a power law but show an exponential decay.

The rank-frequency version of the function P was found to be Zipfian (hence P is Lotkaian) in Adamic and Huberman (2002). There, some consequences of this fact are given. The power-type of the function k6P(k) implies the large degree of skewness meaning that many nodes have few (say just one or two) connections while a few nodes have many connections. This is judged in Adamic and Huberman (2002) as a two edged sword as far as resilience of the network is concerned: if a node fails at random it is most likely one with very few connections, hence its failure will not affect the performance of the network overall. However, if one targets a few (or even one) of the high degree nodes, their removal would require rerouting through longer and less optimal paths. If a sufficient number of high degree nodes are removed, the network itself can become fragmented without a way to communicate from one location to another.

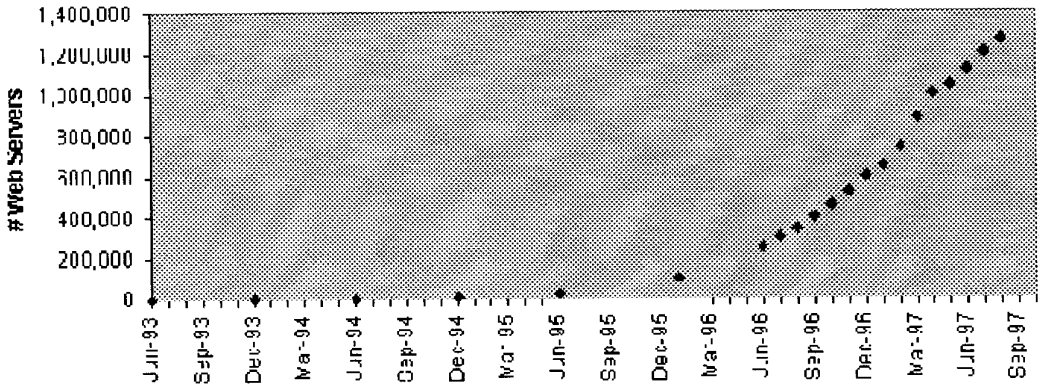
Further examples: in Aida, Takahashi and Abe (1998) and references therein it is reported that het number of URLs which are accessed n times (in a certain time period) is a classical Lotka law with exponent =2 (although they do not use the name Lotka - we need more standardisation, cf. pleads of Glänzel (1996) and Rousseau (2002)). Redner (1998) reports on the number of papers with n citations: Lotka's law is found with exponent .3, a high value.

The examples and laws given above deal with so-called 2-dimensional informetrics (where sources are linked to items). Of course, in informetrics, one can also study time evolutions of 1-dimensional phenomena such as growth. The growth of the Internet is - for the time being - exponential, a very classical distribution indeed. In informetrics (and beyond) it is very well known that growth cannot continue to be exponential. Sooner or later an S-shape arrives. This S-shaped can be

modelled via a Gompertz distribution or a logistic distribution (Verhulst curve) - see e.g. Egghe and Rao (1992a).

I have been looking for S-shapes in graphs on Internet growth. Three years ago I thought I had found one in the growth of the number of web servers (i.e. the number of computers that offer web sites on the net) - see Fig. 3a, found in Netgrowth (1998).

Fig. 3a
 # web servers versus time



However, although Netgrowth (1998) is still active but not updated, I could find the following (spectacular) graph in Adamic and Huberman (2001) Fig. 3b. The small "dent" in 1997 is still there but it is clear that this graph is an example of an exponential growth and not of an S-shaped growth!

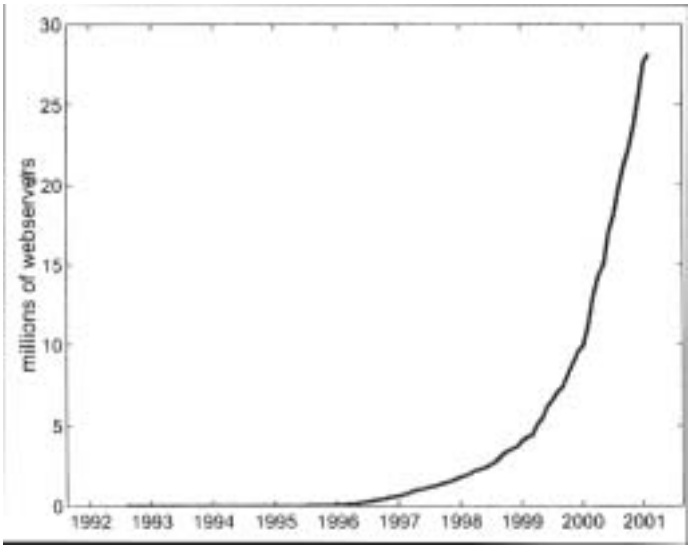


Fig. 3b # webservers versus time (update)

Also all the other growth curves I found are purely exponential. This is illustrated by Fig. 4 on the number of hosts (up to beginning 2002 - one of the most recent data that could be found at the time of the updating of this text (September 2002)). A "host" is any computer system connected to the Internet which has an IP address associated with it (see Mc Murdo (1996)). The graph was found in Internet Software Consortium (2002). In total there are about 160.10^6 hosts, mid 2002. Since january 1993 there has been a multiplication by 1.5, every year. Hence the growth rate is 1.5. Note: the last data point (of mid 2002) "seems" to indicate the start of an S-shape (but remember Figs. 3a-3b!!).

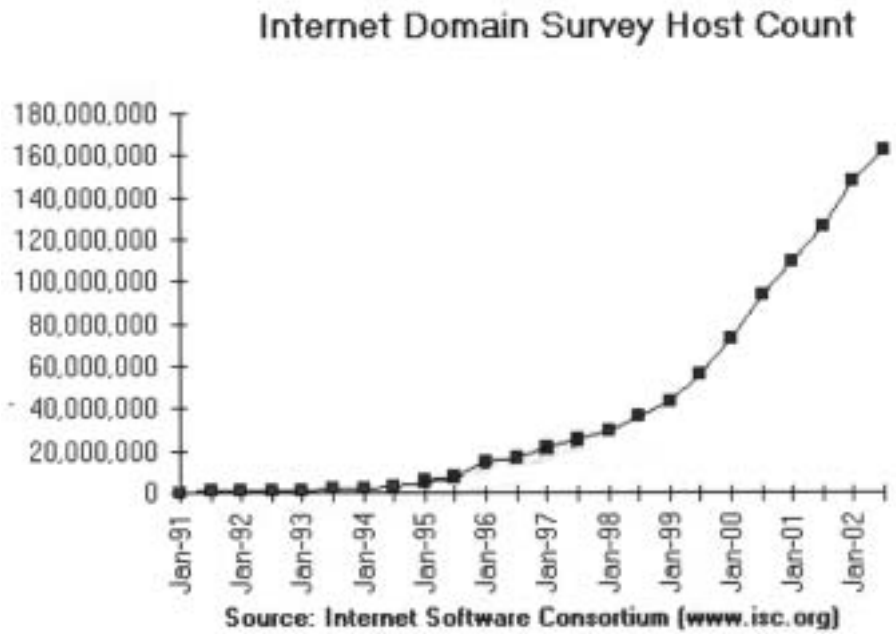


Fig. 4 # hosts versus time

From the above it seems that - although not easy - it is possible to model the growth of WWW or of Internet. In classical informetrics a "dual" companion of growth is ageing for the simple reason that the same techniques that apply to study growth (e.g. growth rate) can be used in the study of ageing (e.g. ageing rate). We will come back to this issue later on but we can already mention here that measuring ageing of the Internet is - conceptually as well as in practise - a far more complex task.

II.3 Complexity in informetrics

We can report here on a recent finding (by the author) of the link between informetrics (IPPs) and self-similar fractal theory. Our findings, however, are based on an old result of Naranan (1970) which goes as follows. Suppose that

(i) The number of sources grows exponentially in time t:

$$N(t) = c_1 a_1^t , \quad (4)$$

(ii) The number of items in each source grows exponentially in time,

(iii) The growth rate in (ii) is the same for every source: (ii) and (iii) together imply an fixed exponential function

$$P(t) = c_2 a_2^t \quad (5)$$

for the number of items in each source at time t.

Then this IPP is Lotkaian, i.e. the law of Lotka applies: if f(p) denotes the number of sources with p items, we have

$$f(p) = \frac{C}{p^{\alpha}} \quad (6)$$

where

$$\alpha = 1 + \frac{\ln a_1}{\ln a_2} \quad (7)$$

The link with self-similar fractals is as follows (not noticed in Naranan (1970)). Let us take the example of a triadic Koch curve (see also Egghe and Rousseau (1990)): we start with a line piece and, at each level, we transform each line piece into 4 line pieces with length 1/3 of the original one (see Fig. 5). So we have

(i) **The number of line pieces grows exponentially in time t, here proportional with 4^t**

(ii),(iii) **1/length of each line piece is the same for every line piece and grows exponential in time t, here proportional with 3^t.**

This triadic Koch curve is an example of a proper fractal. This means that if our scale, say, doubles (e.g. an airplane, from where we watch a coastline, halves its height) the length multiplies with more than 2. The fractal dimension is a way to measure this "more than" and is the alternative for measuring lengths which, strictly speaking, is not possible if we do not indicate from which "heigth" we are watching. A

classical result in self-similar fractal theory is that its fractal dimension is given by (for the triadic Koch curve)

$$D = \frac{\ln 4}{\ln 3} = 1.26186 > 1 \qquad \textbf{(8)}$$

Rephrased in terms of informetrics: a (Lotkaian) IPP (consisting of sources and items - we could call it 2-dimensional informetrics) is a self-similar fractal and its fractal dimension is given by the logarithm of the growth rate of the sources, divided by the logarithm of the growth rate of the items

$$D = \frac{\ln a_1}{\ln a_2} \qquad \textbf{(9)}$$

(which can be > or < 1). Hence, the exponent in Lotka's law satisfies the important relation:

$$= 1 + D \qquad \textbf{(10)}$$

This result was earlier seen by Mandelbrot but only in the context of (artificial) random texts (hence in linguistics).

That a power law (law of Lotka) describes self-similarity of informetrics is also easily seen as follows: suppose we change scales in (6), say changing p into Bp. Then (6) becomes $C|(Bp)^{\alpha} = \frac{C|B^{\alpha}}{p^{\alpha}}$**, hence the same power law. This property is also called "the scale-free" property. This means e.g. that if one were to look at the distribution of site sizes for one arbitrary range, say sites that have between 1,000 and 2,000 pages, it would look the same as that for a different size range, say from 10 to 100 pages. In other words, zooming in or out in the scale at which on studies the web, one keeps obtaining the same result, just as in the case of the Koch curve (Fig. 5) (Huberman (2001)).**

Examples of fractal dimensions: D=1 for a line, D=2 for a surface, D=1.52 for the coastline of Norway (Feder (1988)). In informetrics: D=1 (hence =2) if $a_1 = a_2$**, hence if the growth rate of the sources is the same as the one of the items. Also the higher , the higher D and hence the higher the**

complexity of the IPP. The examples given in subsection II.2 showed high values of Lotka's for the connectivity of social networks. Hence, in view of (10), their fractal dimension (hence their complexity) is high.

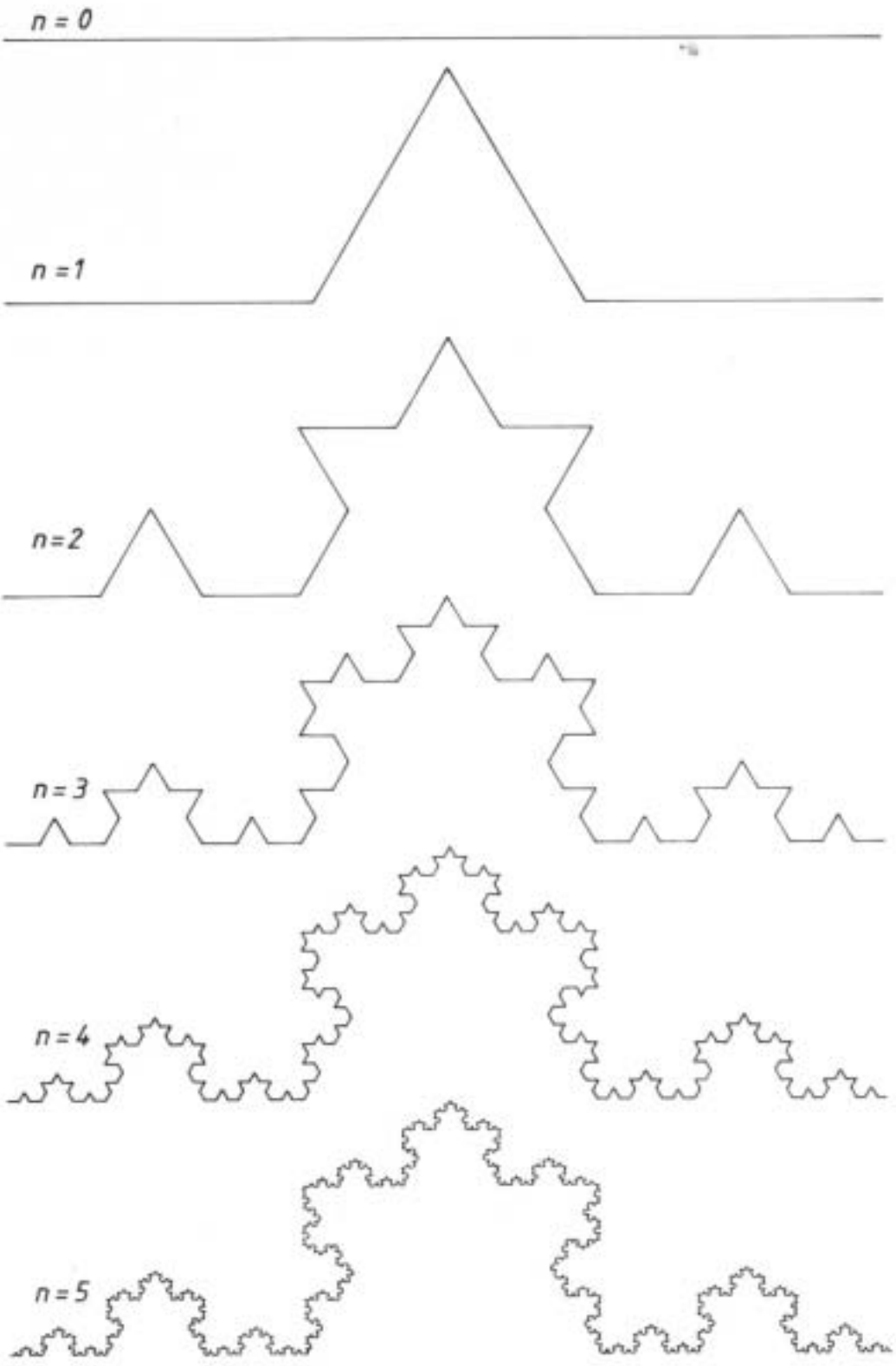


Fig. 5 Construction of the triadic Koch curve

III. Internet and citation analysis.

One of the largest parts of informetrics is scientometrics. Some scientometricians even refuse to consider scientometrics as a part of informetrics. This is the reason for the name of the ISSI society (International Society of Scientometrics and Informetrics) founded in 1993 in Berlin during the third international conference. The discussion on informetrics versus scientometrics is not the topic of this paper, however.

Basic in scientometrical analysis are the references given in books and articles. The web analogue of references is generally accepted to be the clickable buttons in a web page (also called hyperlinks) - see Almind and Ingwersen (1997), CLEVER (1999), Rousseau (1997) or Aguillo (1998). But hyperlinks are very different from references. From a conceptual point of view, citing is a one way link (in mathematical terms a digraph = directed graph) since if article B is cited in article A, A will not be cited in B (some anomalies do exist due to publication delays and the existence of invisible colleges) but at least conceptually it is correct : giving a reference is using older work, i.e. work from the past - hence this relation cannot be reversed. In the framework of hyperlinks, however, web site A can give a hyperlink to web site B and vice-versa. Here the network can allow for bidirectional links.

This conceptual difference originates from the fact that web sites can be updated and hence that a notion such as "publication time" is fuzzy, if at all existent. Almind and Ingwersen (1997) talk in this connection about "real time". And this jeopardises hyperlinks as tools to measure ageing, in contrast with references or citations : they are basic for all ageing studies, see e.g. Egghe and Rao (1992b). Hyperlinks point to URLs and it is well-known that updating URL-addresses is a tough task. For this reason in 1997 a group of American publishers at the Frankfurter Buch Messe proposed the so-called DOIs (Digital Object Identifier) which are unchangeable codes for digital objects, such as (parts of) web sites, an electronic article or even a graph. Their DOIs do not change even if the

URL changes. DOIs are the electronic version of ISBNs and ISSNs and are also created to protect intellectual property. For more information on DOIs, see Simmonds (1997) or the DOI-web site : www.doi.org. One can hope that DOIs will contain some time-element (e.g. year or month that a DOI is given). In this way hyperlinks (to DOIs) could be used in time or ageing studies. It does not seem to incorporate control characters (as is the case in ISSNs and ISBNs), which, I think, is a serious drawback. The DOI-management is now in the hands of the International DOI Foundation (IDF) (Genève, Oxford and Washington D.C.).

Referring again to the big difference between hyperlinks and references, we are not convinced that the definition of Web-impact-factor (WIF) given by Ingwersen (1998) is the right translation of the classical IF to the environment of the WWW. His definition goes as follows : "the WIF is calculated as the sum of the number of web pages pointing to a given country or web site divided by the number of pages found in that country or web site, at a given point in time" (see Ingwersen (1998), p. 237). In fact Ingwersen himself points out some conceptual differences between WIF and IF. One example is that only the number of linked sites counts and not the number of linked pages. Compared to the classical case this would mean that citations in article(s) in a journal A to, say, 2 articles in the same journal B is counted as one citation, which clearly is not the case for the classical IFs. How to do better than Ingwersen is, however, still an open problem.

An application in which the role of hyperlinks is similar to the one of references is given in CLEVER (1999) in the context of information retrieval. We will discuss this in the next section (amongst other things).

IV. Information retrieval in the Internet and its quantitative evaluation.

IV.1 Information retrieval in the Internet.

Internet and more in particular WWW is the world's richest source of information but at the same time it is increasingly difficult to retrieve the right information from this source. This in combination with the fact that more and more untrained people want to retrieve information makes the study of IR in WWW a real challenge (until some years ago the only people that used IR in electronic databases were professional librarians searching in field specific files such as Chemical Abstracts, Inspec, ..., usually collected together by hosts such as DIALOG).

Until the creation of the Internet, IR tools were rather basic, using techniques of Boolean searching (AND, OR, NOT) or of word proximity based on an inverted file structure of key words. Of course we must admit that research in IR has been dealing with more advanced techniques long before the Internet existed but practical implementations of the results of this research have been exceptions. Examples of such research are probabilistic IR and IR based on clustering. We will not discuss these topics in full detail (see for this e.g. Salton and Mc Gill (1987)) but basically all these techniques are based on a vector representation of documents and queries. These vectors have a 1 on coordinate i if term i is present in the query or the document and a 0 if not (more general weights $\in [0,1]$ can be assigned but we will not go into this). In this setting, queries and documents have similar representations and the former can be replaced by the latter or vice-versa. This is one aspect of duality in IR. For more on this we refer to Egghe and Rousseau (1997).

In probabilistic IR one calculates, based on samples, $P(rel|d)$, the probability that a document d is relevant w.r.t. a given query. Alternatively one can apply a matching function between a document d and

a query q. Such a matching function calculates the "degree" of similarity between d and q. An example of this is the Salton cosine formula (see e.g. Salton and Mc Gill (1987)). Other measures (such as the Jaccard or Dice index - see further) exist.

Both techniques yield numbers with the property that the higher they are, the better. Hence documents can be ranked in decreasing order of these numbers. This is better than in conventional IR where one simply presents a set of retrieved documents (hence a query of the form "cat or dog or mouse" equally retrieves documents that deal with these 3 topics or that only deal with one of these topics). Otherwise said, one does not present a set of documents but a ranked list. This is exactly how web browsers present the search results. Of course the above description of ranked output is just a first indication of how things work. Different browsers use different ranking techniques and their exact form is kept a secret ! The same goes for the necessary indexing technique but in any case browsers use automatic indexing techniques for documents (such as idf = inverse document frequency, discriminative value, entropy value and so on - see Salton and Mc Gill (1987)) as a basic tool.

The creation of the Internet and WWW in particular have boosted these advanced techniques of indexing and retrieval into practical everyday use. Yet the problem of IR in Internet is not solved. We still face the problem of selecting the right documents (web pages) from a (usually) very large list. This problem gets worse every day (cf. the above mentioned increase every year - see Fig. 3). This is where cluster IR comes into action. It is still experimental - see the experiments described in CLEVER (1999).

The basic idea is the following (although not exactly followed in CLEVER (1999) but we will come to this further on) : similarities as described above can also be used between two documents d and d_N (hence replacing the query q above by another document d_N). Based on these similarities one can cluster documents, using a technique from multivariate statistics. In this way one forms groups of "similar" documents which is important

knowledge in IR, and these groups are independent from the used query. In addition to this, for each group, one can point out the most central, authorative document(s) and this can be a solution for the large number of ranked documents as a result of an IR process.
The above described technique is applicable in any documentary system. In WWW one can perform even better by modifying the above technique, using the hyperlinks in the web sites (these hyperlinks are called "one of the Web's most precious resources" - see CLEVER (1999), p.49). Here the hyperlinks replace key words in the described technique above. The web sites that are central or occupy an authorative place in a cluster are selected and only these are retrieved. This technique is especially interesting for broad topics but as said above, the larger the web the more "broad" the topics are !

IV.2 Quantitative evaluation of IR in the Internet.

This subsection deals with the many problems of evaluation of IR in the framework of the Internet (say in WWW). Classical techniques are well-known : the evaluation measures are precision P and recall R. Precision is the fraction of the retrieved documents that are relevant. Recall is the fraction of the relevant documents that are retrieved.

In other words, denoting by A the set of retrieved documents and by B the set of relevant documents, we have

$$P = \frac{|A \cap B|}{|A|} \qquad (11)$$

$$R = \frac{|A \cap B|}{|B|}, \qquad (12)$$

where |.| denotes "number of elements (documents) in".

A single measure of IR performance is the harmonic average between P and R,

$$D = \frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2|A \cap B|}{|A|+|B|}, \qquad \textbf{(13)}$$

also known as Dice's measure.
D has the advantage that [D high ⇒ P and R high], which is what we prefer. Of course R and P are always in [0,1] but the closer they are to 1 the better. In practise, however, one experiences that P high causes R to be low and hence a high R causes P to be low. Otherwise stated R is a decreasing function of P (or P is a decreasing function of R) and there are techniques to construct such R-P-curves for an IR-system (see e.g. Salton and Mc Gill (1987)).

We must underline, however, that even in conventional documentary systems, there is always a problem in determining R (P is known from what one retrieves) : indeed, we do not know |B|, the total number of relevant documents (I assume here that the decision on whether a document is relevant or not can be taken upon inspection of the document by the user). A classical technique to determine (a confidence interval for) R is by sampling, analogue to the technique of determining the number of rats in a city ! However this is not part of the every day life of an IR-searcher !

Going to WWW where one performs a search via a web browser (such as Alta Vista - one of the best) we face additional problems. As stated above, we do not even have a set of retrieved documents (A). In answer to a query we obtain a (usually long) ranked list of documents (web sites) of which we examine a certain number X by scrolling from the top of the list to the X^{th} site. Usually X«|A| (here |A| substitutes for the total number of "retrieved" sites - they are not really retrieved as explained above but its length is always given by the system). The way we evaluate such an IR-performance is by calculating the so-called

first - X - Precision **(14)**

(cf. Leighton and Srivastava (1999) and references therein), i.e. the precision obtained in the first X retrieved documents.

Alternatively one can consider the

$$\text{first - Y - Precision} \qquad (15)$$

for every $Y \in \{1,2,\ldots,X\}$.

Of course, in one type of search, (14) suffices : in the case the searcher is only interested in a few (say a=1 or 2) pertinent (relevant) sites, the searcher's happiness (satisfaction) is perfectly measured by (14), where X is the rank of the a^{th} relevant site. An example of such a search is given by a person who wants touristical information on a city or a country he/she wants to visit : this person does not want all the information but just a few pertinent ones.

Bar-Ilan (1998) is an exceptional study where one has dealt with R and P in several search engines (and one studies also overlap amongst them). General conclusion here : P high, R low, overlap low !

We did not go into other evaluation aspects of browsers ; they are more qualitative in nature and compare different features of different browsers. Updated information can be found on the site

http://searchenginewatch.com

<u>We want to close this paper by giving a solution to one of the "new" problems in WWW, namely on extending the "classical" measures R, P, Dice, Jaccard, Cosine,... to ordered sets of documents</u>. As pointed out in subsection IV.1, one does not receive a set of documents as the result of an IR-search but a ranked set of documents. Usually this set is totally ordered : for any two documents d, d' one has that d is above d' or vice-versa. It is important, however, to study also the more general case of a chain as in Fig. 6.

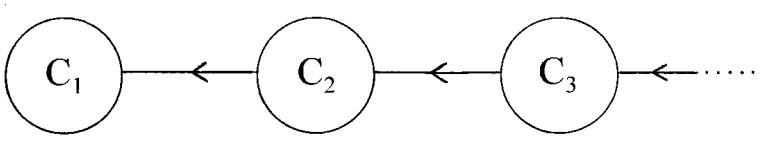


Fig. 6 Partially ordered set of documents.

Here C_1, C_2, C_3, ... are unordered sets of documents (i.e. there is no priority of retrieval between the documents in a set C_i) but all documents in C_1 are retrieved before all documents in C_2 ; all documents in C_2 are retrieved before all documents in C_3, and so on. An example of application of this is given by the example in subsection IV.1 : IR on "cat or dog or mouse". C_1 then contains the documents that contain all 3 key words cat and dog and mouse. Certainly this is the most important set of documents and these documents should be retrieved first. Of course, by the query itself, there is no priority ranking between the documents in C_1. C_2 is the set of documents that contain 2 key words of the given 3. Also this set is unordered but every document in C_2 should be retrieved before every document in C_3, the set of documents that contain only 1 key word of the given 3. Finally C_4 is ranked last and contains the documents that have none of the key words "cat or dog or mouse" in their indexing.

This general situation also comprises the well-known and important cases :

1. **Total order : every C_i is a singleton : $|C_i|=1$, for all i**
2. **Not ordered case : the chain consists of one set : C_1.**

In Michel (2001), Egghe and Michel (2000a,b) one has studied the problem : is it possible to define good similarity measures for ordered sets as in Fig. 6, based on the "classical" similarity measures such as Jaccard, Dice,

Cosine, R or P on unordered sets ? What good measures are (in the case of unordered or ordered sets) will be explained below. First we recall the definitions of the measures mentioned above. The ones of P, R and D have already been given in (11), (12) and (13).

Jaccard's measure is

$$J = \frac{|A \cap B|}{|A \cup B|} \qquad \textbf{(16)}$$

Cosine measure is

$$Cos = \sqrt{RP} = \frac{|A \cap B|}{\sqrt{|A||B|}} . \qquad \textbf{(17)}$$

In Egghe and Michel (2000a,b) other measures appear. Several measures are also described in Boyce, Meadow and Kraft (1995), Grossman and Frieder (1998), Losee (1998), Tague-Sutcliffe (1995), Van Rijsbergen (1979) and of course also in Salton and Mc Gill (1987).

These measures have the following properties (denote F for one of the above similarity measures)

(F_1) $0 \leq F(A,B) \leq 1$

(F_2) $F(A,B) = 1 \Rightarrow A = B$ **(OK for D, J, Cos)**

OR, the weaker

($F_2^{\mathbb{N}}$) $F(A,B) = 1 \Rightarrow A \supseteq B$ **or** $B \supseteq A$ **(OK for R, P)**

(F_3) $F(A,B) = 0 \Rightarrow A \cap B =$

(F_4) **If the denominator of F is constant then F is an increasing function of** $|A \cap B|$.

A measure F satisfying (F_1), (F_2), (F_3), (F_4) is called a good strong similarity measure ; a measure F satisfying (F_1), ($F_2^{\mathbb{N}}$), (F_3), (F_4) is called a good weak

similarity measure (note indeed that $(F_2) \Upsilon (F_2^N)$). In Egghe and Michel (2000a) we managed to construct good strong measures of similarity for ordered sets ; in Egghe and Michel (2000b) we used methods of fuzzy sets to construct good weak measures of similarity for ordered sets. Here we will (briefly) explain how the first type of measures can be constructed.

Let C, C^N be two chains as described above : $C=(C_1, C_2, C_3, ...)$, denoted (C_i), $C^N=(C^N_1, C^N_2, C^N_3, ...)$, denoted (C^N_j). We require the following properties for good strong measures Q of similarity on ordered sets C, C^N :

(Q_0) **If C and C^N reduce to the unordered case, Q must be a good strong similarity measure for unordered sets (i.e. satisfy (F_1), (F_2), (F_3), (F_4)).**

(Q_1) **$0 \# Q(C,C^N) \# 1$, for all chains C, C^N**

(Q_2) **$Q(C,C^N) = 1 \rceil C=C^N$ and no C_i or C^N_j is empty**

(Q_3) **$Q(C,C^N) = 0 \rceil C \perp C^N = \emptyset$. Here we define**

$$C \perp C^N = \left(\bigcup_{i=1}^{4} C_i \right) \perp \left(\bigcup_{j=1}^{4} C^N_j \right)$$

These are natural requirements, "mirrored" from the unordered case. Typically for the ordered case are the following two requirements.

(Q_4) **Let $i,j \in \mathbb{N}$, $i \neq j$. Let $C^{(i)}$, $C^{N(j)}$ be ordered sets (chains) such that $C_k \perp C^N_l = \emptyset$, $\forall k,l \in \mathbb{N}$ except for $k=i$ and $l=j$. If we let i and j vary (but not the unordered sets C_i and C^N_j) then $Q(C^{(i)}, C^{N(j)})$ is strictly decreasing in $j>i$ (i fixed) and $i>j$ (j fixed).**

(Q_5) **The same as (Q_4) but now for $i=j$: now $Q(C^{(i)}, C^{N(i)})$ strictly decreases in $i \in \mathbb{N}$.**

These requirements deal with the natural wish to have a smaller impact on the value of Q for sets on a higher rank in the chain. Indeed, the documents in these sets are less used or not used at all.

These requirements are indeed good properties for strong similarity measures but it is mandatory to prove that such measures exist. Our

starting point will be the classical good strong similarity measures on unordered sets, which we mentioned above.

In Egghe and Michel (2000a) we proved the following theorem (we give here a simpler version) :

<u>Theorem</u> : Let F be any strong similarity measure (on unordered sets). Define, for chains C=(C$_i$), C$^{\mathbb{N}}$=(C$^{\mathbb{N}}_j$) :

$$Q(C,C^{\mathbb{N}}) = \sum_{i=1}^{4} \sum_{j=1}^{4} f(F(C_i,C_j^{\mathbb{N}}))n(i,j) \quad , \tag{18}$$

where

$$n(i,j) = \frac{3}{2^i 2^j 2^{|i-j|}} = \frac{3}{4^{\max(i,j)}} \tag{19}$$

and where f is a strictly increasing function such that f(0)=0, f(1)=1, 0≤f≤1 such that

$$\sum_{j=1}^{4} f(F(C_i,C_j^{\mathbb{N}})) \le 1, \quad \forall i \neq 0 \tag{20}$$

$$\sum_{i=1}^{4} f(F(C_i,C_j^{\mathbb{N}})) \le 1, \quad \forall j \neq 0 \tag{21}$$

Then Q is a good strong similarity measure for ordered sets.

The proof is rather long and technical and can be found in Egghe and Michel (2000a). The above theorem has only value if such functions f exist. In Egghe and Michel (2000a) we showed that the following cases work :

- **For Jaccard's J : f(x)=x leading to**

$$Q_J(C,C^{\mathbb{N}}) = \sum_{i=1}^{4} \sum_{j=1}^{4} \frac{|C_i \cap C_j^{\mathbb{N}}|}{|C_i \cup C_j^{\mathbb{N}}|} \frac{3}{4^{\max(i,j)}} \tag{22}$$

- **For Dice's D : $f(x) = \frac{x}{2-x}$, surprisingly leading to**

$$\mathbf{Q_D(C,C^{\mathbb{N}}) = Q_J(C,C^{\mathbb{N}})} \tag{23}$$

S **For Cosine : $f(x)=x^2$ leading to**

$$Q_{Cos}(C,C^{\mathbb{N}}) = \sum_{i=1}^{4}\sum_{j=1}^{4} \frac{|C_i \cap C_j^{\mathbb{N}}|^2}{|C_i||C_j^{\mathbb{N}}|} \frac{3}{4^{\max(i,j)}} \qquad \textbf{(24)}$$

In Egghe and Michel (2000a) other measures F and corresponding Q_F are studied. Egghe and Michel (2000b) then deals e.g. with the case of weak similarity measures (e.g. R and P). Via fuzzy set techniques we arrived at the following ordered variants for R and P. With the same notation as above :

$$Q_P(C,C^{\mathbb{N}}) = \frac{\sum_{i=1}^{4}\sum_{j=1}^{4} |C_i \cap C_j^{\mathbb{N}}| \frac{1}{2^{\max(i,j)\&1}}}{\sum_{i=1}^{4} |C_i| \frac{1}{2^{i\&1}}} \qquad \textbf{(25)}$$

and the same for Q_R (interchange C and $C^{\mathbb{N}}$).

References

Adamic L.A. and Huberman B.A. (2001). The Web's hidden order. Communications of the ACM 44(9), 55-60.

Adamic L.A. and Huberman B.A. (2002). Zipf's law and the Internet. Glottometrics 3, 143-150.

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.

Aguillo I.F. (1998). STM information on the Web and the development of new Internet R&D databases and indicators. Online Information 98. Proceedings of the 22nd International Online Meeting, London, 8-10 December 1998, Oxford : Learned Information Europe Ltd., 239-243.

Aida M., Takahashi N. and Abe T. (1998). A proposal of dual Zipfian model for describing HTTP access trends and its application to address cache design. IEICE Trans. Comm. E81-B(7), 1475-1485.

Almind T.C. and Ingwersen P. (1997). Informetric analyses on the world wide web : methodological approaches to "webometrics". Journal of Documentation 53(4), 404-426.

Barabási A.-L. and Albert R. (1999). Emergence of scaling in random networks. Science 286 (October 15, 1999), 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.

Bar-Ilan J. (1998). On the overlap, the precision and estimated recall of search engines. A case study of the query "Erdos". Scientometrics 42(2), 207-228.

Bilke S. and Peterson C. (2001). Topological properties of citation and metabolic networks. Physical Review E, 6403(3), 76-80.

Boudourides M.A., Sigrist B. and Alevizos P. (1999). Webometrics and the self-organization of the European Information Society. http://hyperion.math.upatras.gr/ webometrics/.

Boyce B.R., Meadow C.T. and Kraft D.H. (1995). Measurement in Information Science. Academic Press, New York.

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

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

Egghe L. (1990). The duality of informetric systems with applications to the empirical laws. Journal of Information Science 16(1), 17-27.

Egghe L. and Michel C. (2000a). Strong similarity measures for ordered sets of documents in information retrieval. Information Processing and Management, to appear.

Egghe L. and Michel C. (2000b). Construction of similarity measures for ordered sets of documents using fuzzy set techniques. Information Processing and Management, to appear.

Egghe L. and Ravichandra Rao I.K. (1992a). Classification of growth models based on growth rates and its applications. Scientometrics 25(1), 5-46.

Egghe L. and Ravichandra Rao I.K. (1992b). Citation age data and the obsolescence function : fits and explanations. Information Processing and Management 28(2), 201-217.

Egghe L. and Rousseau R. (1990). Introduction to Informetrics. Elsevier, Amsterdam.

Egghe L. and Rousseau R. (1997). Duality in information retrieval and the hypergeometric distribution. Journal of Documentation 53(5), 488-496.

Erdös P. and Rényi A. (1960). On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. 5, 17-61.

Feder J. (1988). Fractals. Plenum, New York.

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

Grossman D.A. and Frieder O. (1998). Information Retrieval. Algorithms and Heuristics. Kluwer Academic Publishers, Boston.

Huberman B.A. (2001). The laws of the Web. Patterns in the Ecology of Information. The MIT Press, Cambridge, Masssachusetts.

Ingwersen P. (1998). The calculation of web impact factors. Journal of Documentation 54(2), 236-243.

Internet Software Consortium (2002). Internet domain survey. http://www.isc.org/ds/ hosts.html. This site is active and updated.

Jeong H., Tombor B., Albert R., Ottval Z.N. and Barabási A.-L. (2000). The large-scale organization of metabolic networks. Nature 407 (October 5, 2000), 651-654.

Leighton H. Vernon and Srivastava J. (1999). First 20 precision among world wide web search services (search engines). Journal of the American Society for Information Science 50(10), 870-881.

Losee R.M. (1998). Text Retrieval and Filtering. Analytic Models of Performance. Kluwer Academic Publishers, Boston.

Mc Murdo G. (1996). The net by numbers. Journal of Information Science 22(5), 381-390.

Michel C. (2001). Ordered similarity measures taking into account the rank of documents. Information Processing and Management 37(4), 603-622.

Naranan S. (1970). Bradford's law of bibliography of science: an interpretation. Nature 227 (August 8, 1970), 631-632.

Netgrowth (1998). Internet growth charts. http://www.tolearn.net/marketing/netgrowth. htm. This site is still active but not updated.

Nielsen J. (1997). Do websites have increasing returns ? http://www.useit.com/alertbox/ 9704b.html. This site is still active but not updated. For many more Internet studies by Nielsen, see http://www.useit.com/alertbox/

Redner S. (1998). How popular is your paper? An empirical study of the citation distribution. The European Physical Journal B 4(2), 131-134.

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

Rousseau R. (2002). Lack of standardisation in informetric 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.

Salton G. and Mc Gill M.J. (1987). Introduction to modern information retrieval. Mc Graw-Hill, Singapore.

Simmonds A. (1997). The 21st century ISBN. The Bookseller, 5 December 1997, 20-22.

Tague-Sutcliffe J. (1995). Measuring Information. An Information Services Perspective. Academic Press, New York.

Van Rijsbergen C.J. (1979). Information Retrieval. 2nd Edition, Butterworths, London.