New informetric aspects of the Internet: some reflections - many problems

Peer-reviewed author version

EGGHE, Leo (2000) New informetric aspects of the Internet: some reflections - many problems. In: Journal of Information Science, 26(5). p. 329-335.

DOI: 10.1177/016555150002600505
Handle: http://hdl.handle.net/1942/783

NEW INFORMETRIC ASPECTS OF THE INTERNET. SOME REFLECTIONS - MANY PROBLEMS

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 five 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). Lastly, complexity

[1] Permanent address

aspects are discussed. The fractal nature of the Internet is highlighted and a modest attempt to measure it is given.

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 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. thesisses), 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 has worsened 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 <u>time</u> ; in the former "use" is measured by <u>number of times</u> 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. 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 India 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. Internet and (classical) informetric laws

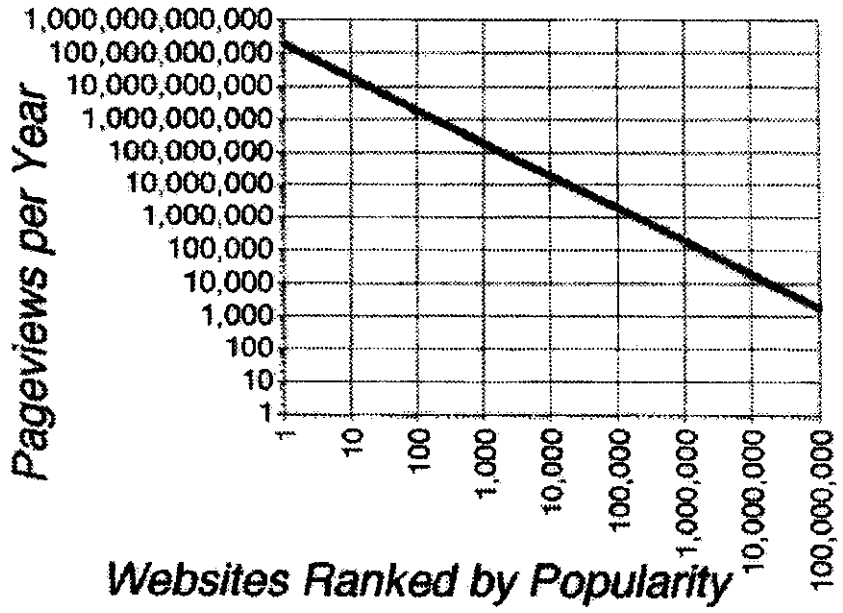
One of the most evident questions that can be asked in this context is : Are the classical informetric laws valid in the Internet ? In other words, are the webometric laws the same as the informetric ones ? This question was also posed by Boudourides, Sigrist and Alevizos (1999) but not at all answered 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 and there is no need to repeat them here. For more information see Egghe (1989, 1990) or the book Egghe and Rousseau (1990).

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. This feature has e.g. been studied in Nielsen (1997) where one produces an acceptable prediction of the number of web sites versus the number of pageviews per year for each site and this for the year 2000. This rank-frequency distribution clearly is Zipfian and an explanation is given using an heuristic SBS (Success Breeds Success) argument (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

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. This also goes for the distribution of domain names (such as .edu, .com, .uk and so on).

The examples 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 but I only 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. 2, found in Netgrowth (1998).

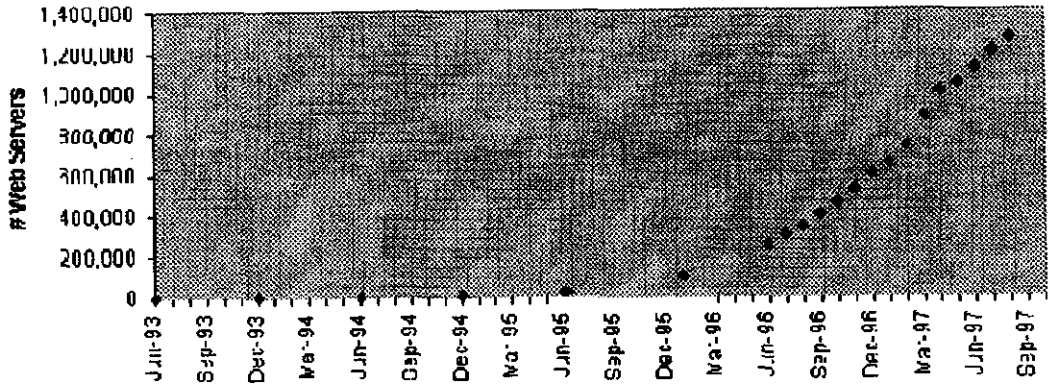


Fig. 2

web servers versus time

All the other growth curves I found are purely exponential. This is illustrated by Fig. 3 on the number of hosts (up to beginning 1999 - one of the most recent data that could be found at the time of the writing of this text (August 1999)). 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 (1999). In total there are about 43.10^6 hosts, beginning 1999. Since january 1993 there has been a multiplication by 2, every year. Hence the growth rate is 2.

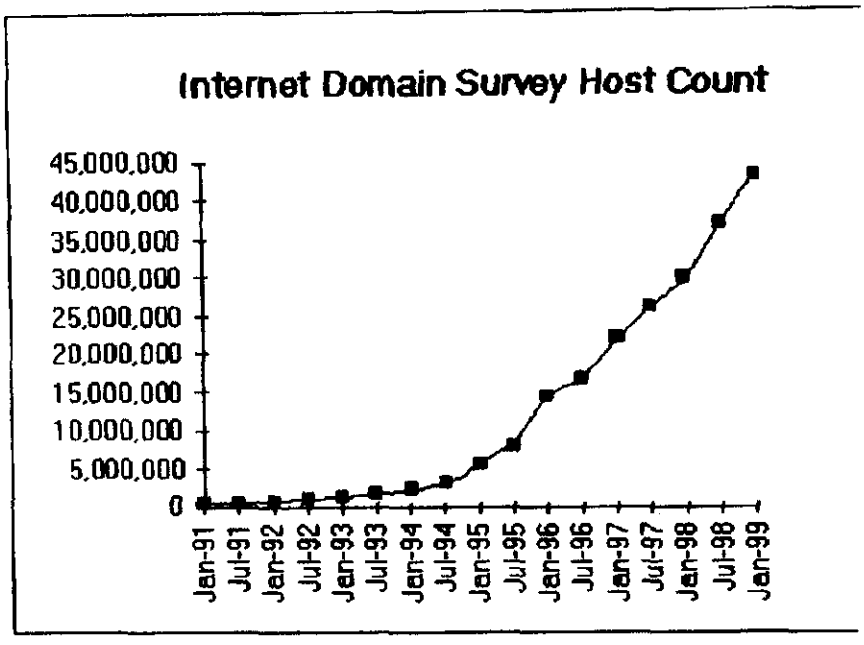


Fig. 3

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.

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 or even a graph. Their DOIs do not change even if the URL changes. DOIs are the electronic version of ISBNs and ISSNs. For more information on DOIs, see Simmonds (1997) or the DOI-web site : www.doi.org. DOIs are not very popular yet and one is still in the experimental phase. 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. The DOI-management is now in the hands of the DOI foundation (Genève 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 link pages counts and not the number of links. Compared to the classical case this would mean that citations 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) 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 doubling in size 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' (hence replacing the query q above by another document d'). 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.

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

$$E = \frac{2}{\frac{1}{P}+\frac{1}{R}}. \qquad (1)$$

E has the advantage that [E high $\Leftrightarrow$ 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 #rel, 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 (ret). 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«#ret (here #ret 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 (2)

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

Alternatively one can consider the

$$\text{first - Y - Precision} \tag{3}$$

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

This is what is available now on IR-evaluation in the Internet. We leave it as an open problem to develop new, dedicated quantitative evaluation techniques for IR in the Internet. In Losee and Paris (1999) one suggests to use ASL, the average search length, being the average position of relevant documents in the ranked list. A "negative" variant of this is the ESL (of Cooper (1968)) being the expected number of (nonrelevant) documents one has to retrieve (e.g. in a ranked list) in order to have a fixed number (k) relevant ones. See also Salton and Mc Gill (1987).

Of course, in one type of search, (2) 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 (2), 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

V. Complexity of the Internet.

That the Internet is a large complex system is an evidency. Here the word "complex" is used in its heuristic sense. There exist exact mathematical tools for measuring complexity, namely in the fractal theory, invented by B. Mandelbrot (1977). Fractal theory is well established in the literature, see e.g. Feder (1988) or the easy introduction in Egghe and Rousseau (1990).

The issue of measuring complexity in information science has not been completely solved yet. Using an argument with self-similar fractals (see e.g. Feder (1988)), Mandelbrot was able to show that for texts consisting of words (note that this is an example of an IPP as considered by Egghe (1989, 1990), if

$$f(r) = \frac{A}{(1+Br)^{\beta}} \tag{4}$$

is the generalisation of Zipf's law denoting the number of times a word on rank r is used (ranks are given according to the number of times the word is used - the mostly used ones gets the smallest ranks), then $1/\beta$ is the fractal dimension of the text. Mandelbrot presupposes lots of simplifications (he assumes that all letters have the same frequency and that they occur independently from each other) but nevertheless one is inclined now to accept that $1/\beta$ is the fractal dimension of any IPP whenever (4) is valid. A rationale is lacking for it, until now.

It is clear from the above that measuring the fractal dimension of the Internet or of WWW is not an easy thing to do. In Egghe (1997) a first attempt has been given, applying Mandelbrot's model for texts. We could prove

Theorem (Egghe) : Let N denote the total number of web pages and let μ be the average number of hyperlinks per page. Then the fractal dimension D of this hypertext system is given by

$$D = \frac{\log N}{\log N + \log\left(\frac{1+\mu}{\mu}\right)} \tag{5}$$

We have no applications of this nor do we have other fractal models for hypertext systems such as WWW. This is definitely a large open problem.

References

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.

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.

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.

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/.

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

Cooper W.S. (1968). Expected search length : A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. Journal of the American Society for Information Science 19(1), 30-41.

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. (1997). Fractal and informetric aspects of hypertext systems. Proceedings of the sixth Conference of the International Society for Scientometrics and Informetrics, Jerusalem (16-19 June, 1997), (B. Peritz and L. Egghe, eds.), 71-79. The Hebrew University of Jerusalem. Reprinted in Scientometrics 40(3), 455-464, 1997.

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.

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

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

Internet Software Consortium (1999). Internet domain survey. http://www.isc.org/dsview.cgi?domainsurvey/hosts.gif

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. and Paris L.A.H. (1999). Measuring search-engine quality and query difficulty : ranking with Target and Freestyle. Journal of the American Society for Information Science 50(10), 882-889.

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

Mandelbrot B. (1977). The fractal geometry of nature. Freeman, New York.

Netgrowth (1997). Internet growth charts. http://tolearn.net/marketing/netgrowth2.html

Nielsen J. (1997). Do websites have increasing returns ? http://www.useit.com/alertbox/9704b.html. This site is still active. For many more Internet studies by Nielsen, see http://www.useit.com/alertbox/

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

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.