Fractal and informetric aspects of hypertext systems

FRACTAL AND INFORMETRIC ASPECTS OF HYPERTEXT SYSTEMS

by

L. EGGHE

LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium(*)
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium
e-mail : legghe@luc.ac.be

ABSTRACT

The paper studies fractal features (such as the fractal dimension) of hypertext systems (such as WWW) and establishes the link with informetric parameters. More concretely, a formula for the fractal dimension in function of the average number of hyperlinks per page is presented and examples are calculated. In general the complexity of these systems is high.

This is also expressed by formulae for the total number of hypertext systems that are possible, given a fixed number of documents.

I. INTRODUCTION

The complexity of hypertext systems such as the World Wide Web (WWW) is a generally accepted fact. We refer here to the intricate "web" of hyperlinks, that can exist between documents and not to the total number of bytes that are available in all the pages of these documents and certainly not to the retrieval tools that are at our disposition. In this connection we, however, feel that a complexity study of information retrieval in these systems would be very interesting. This study is postponed to a later occasion.

(*) Permanent address.

Acknowledgements : The author is grateful to profs. Dr. R. Philips and R. Rousseau for stimulating discussions that improved the present version of this paper.

Key words and phrases : fractal, informetry, hypertext system, WWW, hyperlink.

So in this paper we will restrict our attention to the complexity of hypertext systems as it is revealed to us via the existing hyperlinks that are available in every document. It is clear that - up to now - the notion of complexity is intuitive. It is however not the merit of this paper to formalise this notion. Indeed, the complexity of various aspects of nature has been formalised by B.B. Mandelbrot, the founding father of the so-called fractal theory. In his basic books, (Mandelbrot 1954, 1977) discusses fractal features of various aspects existing in the physical and virtual world.

The appealing way to describe complexity of nature is by considering the problem of measuring distances. For instance : how long is the coast line of Norway? The complexity of this problem is clear if one considers maps of Norway of different scales. One would feel that a multiplication of the scale, say by 2, would lead to a multiplication of the length of the coast line by 2 but this is not the case : it will be more. The degree of this "more" forms the basis of the notion of fractal dimension : the coast line does nog act as a curve of one dimension. Of course it is not so complex that it stretches out over a two dimensional part of a plane (say a part with a strictly positive area) but the "thing" (i.e. the coast line) is a fractal with dimension between 1 and 2. That the complexity of the coast line of Norway is indeed very high is clearly shown by the fact that its fractal dimension D is 1.52. This and more can be found in Feder (1988) and also in Egghe and Rousseau (1990) where the topic of fractal theory is applied to the area of information science.

Also in te work of Mandelbrot a calculation of the fractal dimension of texts is given, where texts are considered as a concatenation of symbols and spaces (symbols are letters, numbers, ...). It is this model that we will apply in this paper. In the case of hypertext systems, however, we will be able to do more than Mandelbrot could do for texts : in the sequel it will be possible to give an explicit formula for the fractal dimension of hypertext systems in function of concrete well-known informetric parameters such as the average number of hyperlinks per document. This parameter is easy to estimate by statistical methods.

The paper also studies dual versions of hypertext systems and deals with the effect of sampling on the determination of the fractal dimension.

The paper closes by mentioning some open problems.

II. THE FRACTAL DIMENSION OF HYPERTEXT SYSTEMS

A hypertext system (HS) can be considered as a string of vertical and horizontal stripes as in figure 1.

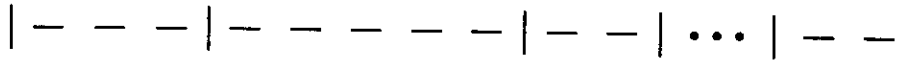

Fig.1 : Symbolic visualization of a hypertext system

Here the vertical stripes represent pages (possibly comprising more than one screen, of course) and the horizontal stripes represent the different hyperlinks (HL), i.e. the clickable parts in these pages. This could be compared to Mandelbrot's model of texts where the vertical stripes denote blanks, being delimiters of words, and where the horizontal stripes denote the letters (or symbols) that form the words.

To show explicitely what we mean by figure 1 we have in this case 3 HLs in the first page, 5 HLS in the second page, 2 HLs in the third page and so on. The order between the pages has no importance in our model (as is also the case in any HS).

Continuing with our HS, denote :

$$n = \text{total number of different HLs} \tag{1}$$

and

$$\varrho = P(\text{a specific HL appears in the HS}). \tag{2}$$

(P = probability)

As in the Mandelbrot model we will assume that ϱ is fixed. This is not so, not here and not in the model of Mandelbrot, but it represents a first approximation. Since there are vertical stripes we have $\varrho < 1/n$. Based on a self-similarity argument (see e.g. (Feder 1988)), we have that the fractal dimension D of such a HS is

$$D = -\frac{\ln n}{\ln \varrho} . \tag{3}$$

It is clear from the above that $0 \leq D \leq 1$ and this is also in accordance with the 1-dimensional visualization of a HS in figure 1. For more on self-similar fractals we refer the reader to (Feder 1988) or to (Egghe and Rousseau 1990).

So far our model and Mandelbrot's are the same. This is also where the derivation of Mandelbrot's model for texts stops. In our case, however, we can do a lot more. Indeed our model of a HS can use the extra property that the total number of different HLs is equal to the total number of pages. In other words :

$$\begin{aligned} n &= \text{total number of different HLs} \\ &= \text{total number of pages.} \end{aligned} \tag{4}$$

Compare this with Mandelbrot's text : the total number of different symbols is certainly not equal to the total number of words!

This fact, typical for HSs, will yield the link with informetrics. The argument is as follows :

$$\begin{aligned} \varrho &= P(\text{a HL appears in the HS}) \\ &= P(\text{a HL appears in the HS} \mid \text{there is a -}).P(\text{-}). \end{aligned}$$

Here P(-) denotes the probability to have a HL and P(...|...) denotes the conditional probability.

$$\rho = \frac{P(-)}{n} \tag{5}$$

Indeed, given the fact that there is a -, and since there are n -'s, supposing they have an equal chance to appear (cf. also the case in Mandelbrot's argument), we have that the conditional probability equals 1/n.
But, typically in HSs, using (4), we have

$$P(-) = \frac{\#-}{\#|+\#-} = \frac{n\mu}{n + n\mu} = \frac{\mu}{1 + \mu} \tag{6}$$

where # denotes : "the total number" and where μ is the average number of HLs per page. Using (5) and (6) we find :

$$\rho = \frac{\mu}{n(1 + \mu)} \quad . \tag{7}$$

(7) in (3) yields :

$$D = \frac{\ln n}{\ln n + \ln\left(\frac{1 + \mu}{\mu}\right)} \quad . \tag{8}$$

We hence have proved the following theorem :

Theorem II.1 :
Let us consider a HS of the form as in Fig.1. Let n denote the total number of pages and μ the average number of HLs per page. Then the fractal dimension D is given by

$$D = \frac{\ln n}{\ln n + \ln\left(\frac{1 + \mu}{\mu}\right)} \quad . \tag{8}$$

Formula (8) represents a link between the fractal theory of HSs and the informetric theory of HSs : indeed, in informetry, pages are sources and the HLs in these pages are the items that are "produced" by these sources. For more on these so-called "Information Production Processes" (IPPs) we refer the reader to (Egghe 1989, 1990) or to (Egghe and Rousseau 1990). So n is the total number of sources and μ is the average number of items per source, hence informetric parameters. Note also that, if $\varphi(j)$ denotes the frequency distribution of the number of sources with j items (in HSs : the number of pages with j HLs), $j \in \mathbb{N}$, we have that

$$\mu = \sum_{j=1}^{n} j\, \varphi(j) \quad . \tag{9}$$

Note also that formula (8) is easy to determine in practise : elementary statistical work, based on the Central Limit Theorem (CLT), allows to determine μ quite accurately from relatively small samples in the HS. This is an advantage since, usually, HSs are very large (cf. the WWW!).

Some properties of the fractal dimension D now follow :

Proposition II.2 :
If the HS is such that each page has at least one HL, then
(i) $\lim_{n\to\infty} D = 1$, independant of the behavior of μ,
(ii) D is a strictly increasing function of μ,
(iii) D has a minimum and a maximum value :

$$D_{min} = \frac{\ln n}{\ln (2n)} \tag{10}$$

$$D_{max} = \frac{\ln n}{\ln \left(\frac{n^2}{n-1}\right)} \quad . \tag{11}$$

The proof of Proposition II.2 is elementary and is left to the reader. It can be obtained from the author, upon request.

Combinatorial Note.

Another (more combinatorial) way of looking at complexity of HSs is by studying the "variations" that are possible in these systems. In this section we will give the number of possible different HSs, but from 4 different points of view (going from high to low level of distinction) :

1. We diversify between the **different** HLs and between the **pages** on which they occur.
2. Only the **number** of different HLs per page is used but we still diversify between the **pages** on which they occur.
3. Only the **number** of different HLs per page counts.
4. Only the **number** of different **fractal dimensions** counts.

Let n be the total number of pages (as in the previous section). Denote

$V_i(n)$ = the number of possibilities in the above cases (i = 1,2,3,4).

In this note we adapt the following (non-controversial) restrictive rules :
(i) The order between the pages is not important.
(ii) The order between the HLs in the pages is not important.
(iii) Each page has at least one HL.
(iv) A page does not give a HL to itself.
(v) Repetition of the same HL on one page is counted as one HL.

In this case we have the following results :

Proposition II.3 :
For every $n \in \mathbf{N}$, $n \geq 2$,

$$V_1(n) = (2^{n-1} - 1)^n \tag{12}$$

$$V_2(n) = (n-1)^n = \overline{P}_n^{n-1} \tag{13}$$

$$V_3(n) = \binom{2n-2}{n} = \overline{C}_n^{n-1} \tag{14}$$

$$V_4(n) = (n-1)^2 \quad . \tag{15}$$

Here $\overline{P}_n^{n-1}$ denotes the number of permutations of n elements selected from n-1 elements with free repetition and $\overline{C}_n^{n-1}$ denotes the number of combinations of n elements selected from n-1 elements with free repetition.
Note that $V_4(n)$ also equals the number of different μ-values that are possible. This follows from proposition II.2(ii).
Note also that, by definition,

$$V_4(n) \leq V_3(n) \leq V_2(n) \leq V_1(n) \ ,$$

$\forall n \in \mathbf{N}$, $n \geq 2$.

The proofs of formulae (12)-(15) are elementary but can be obtained from the author, upon request.

III. SOME REMARKS AND OPEN PROBLEMS

III.1. Dual theory

In Egghe (1989,1990) the notion of duality in IPPs has been explained. By this we mean the interchange of the sources and the items in the informetric study of these processes.

In our framework this boils down to the consideration of a HS as a set of HLs (now the sources) that "produce" pages (now the items) by which we indicate the pages in which these HLs appear. The dual of figure 1 is now

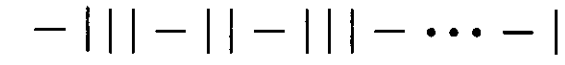

Fig.2 : Dual vision on a HS : HLs that contain pages

Now each different HL(-) is followed by the pages in which it occurs.

There is no essential difference between figures 1 and 2 and hence our theory of section II applies. We now have that

$$D' = \frac{\ln n'}{\ln n' + \ln\left(\frac{1 + \mu'}{\mu'}\right)} \tag{16}$$

where the primes indicate that we deal with the dual situation. We have the following result.

Theorem III.1.1 :
$n = n'$, $\mu = \mu'$ and hence

$$D = D' = \frac{\ln n}{\ln n + \ln\left(\frac{1 + \mu}{\mu}\right)} \; . \tag{17}$$

Since the proof is short, we present it here.

Proof :
By (4),

n' = the number of different HLs
= the number of different pages
= n

Since these numbers form the denominators in the definition of μ and μ' it suffices to prove that their nominators are equal : we have to show that the total number of HLs in the original HS vision equals the total number of pages in the dual vision (hence in both case **with** repetition). This is trivial by definition of the dual HS.

□

Note :
In general IPPs one does not always have that $\mu = \mu'$. This is because (4) is not necessarily valid in other informetric systems (e.g. giving references vs. being cited).

III.2. Samples of HS

Condition (4) expresses a kind of closedness of the HS. Of course, when we perform a search in such a HS, we hence have a subset of the pages appearing in the HS, or otherwise stated, a sample. In this case, (4) is not necessarily valid. Suppose n is our sample size (i.e. the number of pages) and m = the number of different HLs appearing in these pages. Now, most likely, $m \neq n$. An argument as in section II now yields :

$$D_s = \frac{\ln m}{\ln m + \ln\left(\frac{1+\mu_s}{\mu_s}\right)} \qquad (18)$$

where the indices s indicate the sample.

Let us consider the frequency distribution φ (cf. (9)) of the HS, i.e. $\varphi(j)$ = the fraction of pages with j HLs. It remains to be studied what type of distribution φ is (our guess : a discrete approximation of the lognormal distribution, based on previous results - see many references on the lognormal distribution in informetrics) but I think it is fair to assume at least that the Gaussian statistics (hence the CLT) applies. Based on these results we can assume that $\mu \approx \mu_s$ if our search is large enough (i.e. if n is large enough). Hence (18) yields

$$D_s \approx \frac{\ln m}{\ln m + \ln\left(\frac{1+\mu}{\mu}\right)} \quad . \qquad (19)$$

III.3. Problems

1. Determine the frequency distributions φ for the HS and ψ for the dual HS (conjecture : both are lognormal).
2. $\lim_{n\to\infty} D = 1$. This means that for most HSs, $D \approx 1$. Transform D so that more difference is found between different HSs.
3. The fractal arguments given here could be considered as "static" in the sense that they only use the pages and the HLs appearing in them. No IR argument has been used. We are convinced that a fractal theory of HSs is possible adding IR in the ingredients. So, describe the fractal nature of HSs where one considers an IR process.
   This could be depicted as in figure 3.
   This idea has been communicated to me by Philips (1996).
   It is the guess that the fractal dimensions D in these cases will satisfy $D \geq 1$.

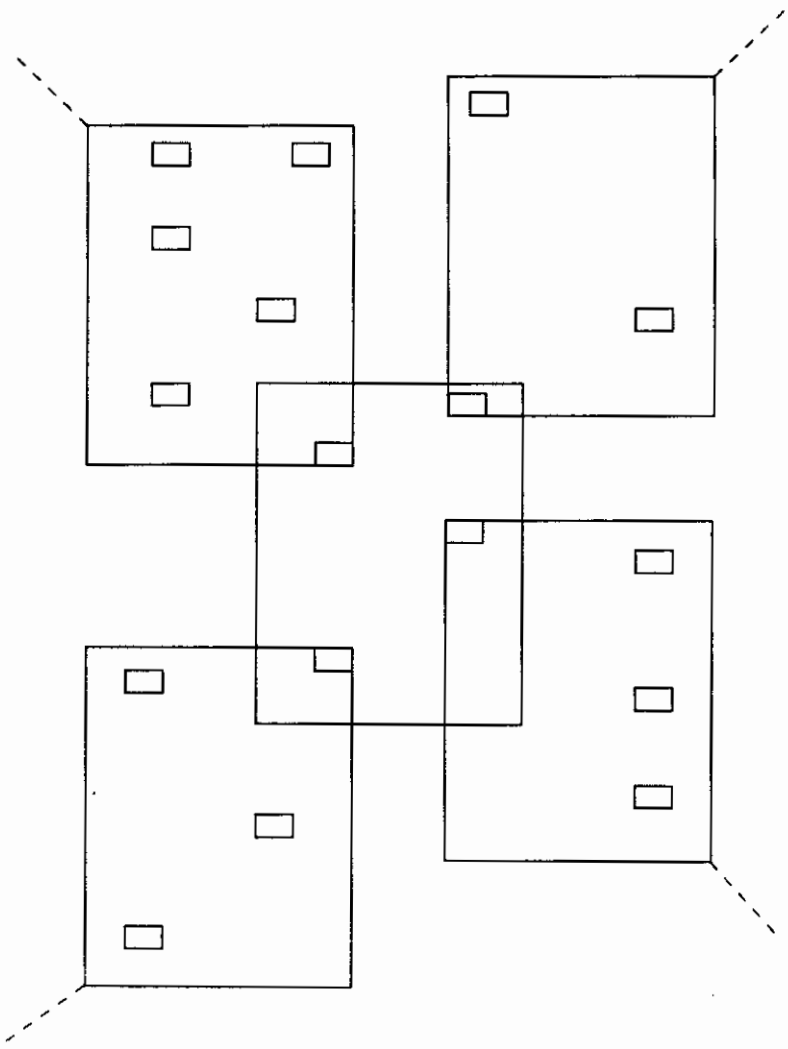

Fig.3 : Possible visualisation of IR in HSs

REFERENCES

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

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

EGGHE, L. (1995)
Extension of the general "success breeds success" principle to the case that items can have multiple sources. In : Michael E.D. Koenig and Abraham Bookstein, eds., *Proceedings of the Fifth Biennial Conference of the International Society for Scientometrics and Informetrics* (Rosary College, River Forest, Il., USA, 1995), 147-156. Learned Information, Inc., Medford, NJ.

EGGHE, L. (1996)
Source-item production laws for the case that items have multiple sources with fractional counting of credits. *Journal of the American Society for Information Science* 47 (10): 730-748.

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

EGGHE, L. and ROUSSEAU, R. (1995)
Generalized success-breeds-success principle leading to time-dependent informetric distributions. *Journal of the American Society for Information Science* 46 (6): 426-445.

EGGHE, L. and ROUSSEAU, R. (1996)
Stochastic processes determined by a general success-breeds-success principle. *Mathematical and Computer Modelling* 23 (4): 93-104.

FEDER, J. (1980)
Fractals. Plenum, New York.

MANDELBROT, B. (1954)
Structure formelle des textes et communications.
Word 10: 1-27.

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

PHILIPS, R. (1996)
Oral communication.