International Journal of Scientometrics, Informetrics and Bibliometrics
ISSN 1137-5019

December 2, 2003

VOLUME 1 (1997): ISSUE 1. PAPER 1

Sitations: an exploratory study

Ronald Rousseau
KHBO, Faculty of Industrial Sciences and Technology
Zeedijk 101, 8400 Oostende, Belgium
Ronald.Rousseau@kh.khbo.be

Abstract

We investigate the distribution of domain names and the distribution of links between web sites. It is shown that the Lotka function provides an adequate description. The percentage of self-sitations is also determined.

Keywords

web sites, sitations, self-sitations, Lotka's distribution, AltaVista

Introduction

The aim of this study is to perform a preliminary study of links between sites on the Internet. We are, however, not concerned with the number of links that can be found on a given web page, but we are interested in the number of times a given web page is referred to. In terms of classical citation analysis: we do not study references, but citations.

We will use the term 'sitation' to designate this relation between sites on the Internet. The term 'sitation' in the sense of cited sites has been advanced by McKiernan (1996) and has been used, e.g., by Aguillo during the 4S/EASST meeting at Bielefeld (October 1996). Studying this kind of links is conceptually the same as studying citations between published articles. The meaning, however, is probably slightly different.

Although there are numerous reasons for citing (cf. Egghe & Rousseau, 1990) from p. 211 on), it is generally agreed that persuasion is probably the most important reason for a citation (Gilbert, 1977), (Brooks, 1985). Web pages, however, are usually not scientific articles and links are probably more made to inform readers where to find more information about the issues presented or discussed on the page. However, as far as I know, reasons why people link their page to other ones have not been investigated yet.

Our work is related to Egghe's dual theory of hyperlinks (Egghe, 1997) and Bar-Ilan's study of Usenet newsgroups (Bar-Ilan, 1997). More sophisticated informetric modelling of the WWW was done e.g. by Larson (1996), Huberman and Lukose (1997) and Almind and Ingwersen (1997). We will present here another illustration that 'old bibliometrics' is applicable to the 'new' Internet

phenomenon.

Method

On May 14 1997 we did a search in AltaVista – Advanced Queries on

bibliometrics OR informetrics OR scientometrics

This resulted in 343 hits, which we downloaded. Note that AltaVista normally shows only the 200 first hits. By 'normally' we mean in the case of simple queries, or in the case of advanced queries, when a ranking term is given. It is only when using advanced queries and no ranking term is indicated that it is possible to go beyond the 200-item limit. This probably explains the problems Bar-Ilan (1997) encountered, when Alta Vista reported 300 relevant items, but displayed only 200. Although it is probable that we obtained a number of false drops, this does not matter for this investigation. We just wanted a sample of sites and to find out how many other sites referred to them. To obtain these numbers we typed the following query (again in AltaVista – Advanced Queries):

link:address AND NOT (url:address)

e.g.:
link:http://sahara.fsw.leidenuniv.nl/ed/edhmpg.html
AND NOT
(url:http://sahara.fsw.leidenuniv.nl/ed/edhmpg.html)

The 'AND NOT'-part removes links from a site to itself. However, in the cases we have checked (we did not do every search twice) this was usually not necessary. These searches were done, manually, during the month of July.

There was one problem though: addresses ending on `:number`, such as

http://www.williams.edu:803/BSC/….

were not recognised by AltaVista's search machine and were rejected as 'bad queries'. The help desk from AltaVista told us that to eliminate the problem one should remove the "http://" part from the query formulation, which indeed worked. Note that an Internet server provides different services such as ftp, telnet, rlogin, connection with the WWW etc. Normally the web server uses port 80. If this is not the case, a number placed at the end of the first part of the address (803 in the above example) indicates this.

Further, sitations of these 343 hits and (for most of the queries) self-sitations were determined. Here a self-sitation is defined as a sitation with the same first component in the address. When searching for all sitations of e.g.

alcazaba.unex.es/english/plandy_e.html

the result

alcazaba.unex.es/english/licence_e.html

is a self-sitation, while

alcazaba.es/biograph.htm

is not (the examples are fictitious).

Results

We first present the search result of the original query *bibliometrics OR informetrics OR scientometrics*. In particular, Table 1 shows that the domain indicators follow a Lotka function. For the meaning of these symbols we refer to, e.g. (ISO 3166, 1997). Here 'number' refers to a BITNET address or any other address, of which the first part consists completely of numbers, such as in **http://164.11.100.12/siv/htm/docs/wwwvrlib.htm**.

The influence of the Marseille team (fr) and of the Danish School of Librarianship and the COLIS conference (dk) is clearly visible. Further, 'nl' means mainly Leiden (Van Raan's team) and Amsterdam (Leydesdorff's team), while the 'be'-sites are just different versions of Leo Egghe's and Ronald Rousseau's publication lists. There were also seven web pages from Spain.

These 343 data points follow the ubiquitous **Lotka distribution**. Indeed Table 1 fits the function (using a maximum likelihood approach (Nicholls, 1987), (Rousseau,1993)):

$$f(y) = \frac{0.4055}{y^{1.54}}$$

where f(y) denotes the relative number of domain indicators that occurred y times. The fit is accepted by a **Kolmogorov-Smirnov** test (5% level).

Table 1. Frequencies of domain indicators of the original query

rank	domain	frequency
1	edu	100
2	fr	33
3	com	29
4	org	22
5	ca	20
6	de	19
	uk	19
8	nl	13
9	au	9
10	es	7
	'number'	7
12	jp	6
	dk	6
14	be	5
15	kr	4
	at	4

	se	4
	my	4
19	il	3
	ch	3
	cl	3
	pl	3
	nz	3
24	us	2
	hr	2
	cz	2
27	11 domains: gov,mx,it,su, ro,hk,net,br,si,fi,in	1

The average size of these web pages was 73.6 K. Note the difference here with the sizes reported in e.g. (Almind & Ingwersen, 1997) or (Woodruff et al., 1996). These authors report pure text sizes after all markup (HTML codes) had been extracted. Our average size is as given by AltaVista and includes all extra codes. The age distribution of the web pages we downloaded was as follows:

129 dated 1997
167 dated 1996
40 dated 1995
6 from 1994
1 from 1993

showing the recent character of web documents.

Sitations

Table 2 presents the distribution of sitations of these 343 hits. This means we have counted all web pages that have at least one link to any of these 343 'informetric' sites. It is clear that there was one extremely popular site, namely **www.w3.org/pub/DataSources/bySubject/** , the subject catalogue of the WWW Virtual library. Most sites, however, were 'unsited'.

Table 2: Frequency distribution of sitations

Number of sitations	Number of sites with this number of sitations
3631	1
89	1
52	1
40	1
37	1
27	1
26	1
11	1
10	1
7	3
6	1
5	6
4	3
3	9
2	19
1	44
0	235

We were not able to fit one of the classical statistical distributions to this set of data. If, however, we assume an implicit link from each site to itself (as done e.g. in (Egghe, 1997), then again Lotka's function provides an excellent fit. More precisely, the function

$$f(y) = \frac{0.7096}{y^{2.345}}$$

where, f(y) denotes the number of sites with y sitations, yields an almost perfect fit (in the case we leave the most-cited web site out; otherwise the parameters of Lotka's function are: 2.295 and 0.6967). This means that the original data can be described by the distribution function

$$f(y) = \frac{0.7096}{(y+1)^{2.345}}$$

Self-sitations are distributed as follows:

Sitations	Self-sitations
1	of the 44, 24 were self-sitations
2	of the 19 cases 8 had no self-sitations, 6 had one and 5 had two self-sitations
3	of the 9 cases, 3 had no self-sitations, while there were two for the other three cases (1,2 and 3 self-sitations)
4	of the three cases there were two with no self-sitations and one with four
5	3 with no self-sitations, one with one, and two with five
6	this site had no self-sitations

7	one with seven self-citations, one with two and one with one
10	this site had seven self-sitations
11	this site had no self-sitations.

We moreover checked the sites with 27, 37 and 40 sitations. They had respectively four, seven and no self-sitations. The case of 3634 sitations had no self-sitations among the first ten sitations. All in all, of the 313 sitations checked, 95, or 30%, were self-sitations.

Conclusions and suggestions for further research

Finding out the distribution of subjects, scientific or other, over domains is a new area for informetric research. In case of the 'subject' *bibliometrics-scientometrics-informetrics*, we found that the distribution followed Lotka's function, with the domain name *edu* as a clear leader. The word *informetrics*, however, is not only a scientific field, but also the name of more than one company.

As the contents of sites on the Internet are very volatile, an investigation such as this cannot be repeated and provide exactly the same results. This means that people who do this kind of studies must always be prepared to forward their data to other scientists in case they want to check the data, reinvestigate the statistical analysis or reuse the data set for other purposes. On the other hand it might also be interesting to repeat this kind of measurements at different times, to find out how volatile the Internet really is, and to see if statistical regularities persist.

We admit this piece of research is very preliminary. In more elaborate investigations robots should be written to collect the data, hence eliminating the manual work in data collection.

Finally, as mentioned in the introduction, it would be interesting to find out why people link their pages to other ones, and to discuss the differences between 'websiting' and citations in scientific articles.

Acknowledgements. We thank Brendan Rousseau (Fontys Hogeschool, Eindhoven, Netherlands) and William De Cat (KHBO, Oostende, Belgium and Leeds Metropolitan University, UK) for help during our research. We also thank Peter Ingwersen (Royal School of Library and Information Science, Copenhagen, Denmark) for stimulating conversations about 'Webometrics'.

References

Almind,T.C. and Ingwersen,P. *Informetric analyses on the World Wide Web: methodological approaches to 'Webometrics'*. **Journal of Documentation** (1997) 53(4), 404-426.

Bar-Ilan, J. *The "mad cow disease", Usenet newsgroups and bibliometric laws.* **Scientometrics** (1997) 39(1), 29-55.

Brooks, T.A*. Private acts and public objects: an investigation of citer motivations*. **Journal of the American Society for Information Science** (1985) 36, 223-229.

Egghe, L. *Fractal and informetric aspects of hypertext systems*. **Proceedings of the sixth conference of the International Society for Scientometrics and Informetrics, Jerusalem, Israel, (Peritz & Egghe, eds.),** (1997) 71-79.

Egghe, L. and Rousseau, R. **Introduction to Informetrics. Quantitative**

methods in library, documentation and information science. Elsevier: Amsterdam, 1990.

Gilbert, G.N. *Referencing as persuasion*. **Social Studies of Science** (1977) 7, 113-122.

Huberman, B.A. and Lukose, R.M. *Social dilemmas and Internet congestion*. **Science** (1997) 277, 535-537.

ISO 3166. *Codes from ISO 3166*, updated by the RIPE network Coordination Centre. < **http://hike1.hike.te.chiba-u.ac.jp/ikeda/ISO/iso3166.txt** > (Jan. 7, 1997).

Larson, R. R. *Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of cyberspace*, (1996). Available at: <**http://sherlock.berkeley.edu/asis96/asis96.html**>. Also in: **Global complexity: information, chaos and control. Proceedings of the 59th annual meeting of the ASIS (Steve Hardin, ed.)** 1996.

McKiernan, G. CitedSites(sm): Citation Indexing of Web resources. **http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm** (1996)

Nicholls, P.T. *Estimation of Zipf parameters*. **Journal of the American Society for Information Science**, (1987) 38, 443-445.

Rousseau, R. *A table for estimating the exponent in Lotka's law*. **Journal of Documentation** (1993) 49, 409-412.

Woodruff, A., Aoki,P.M., Brewer,E., Hauthier,P. and Rowe,L.A. *An investigation of documents from the World Wide Web.* **Fifth International World Wide Web Conference, Paris, France, May 6-10, 1996. http://www5conf.inria.fr/fich_html/papers/P7/Overview.html**

Recieved 13/October/97
Accepted 20/Novembre/97