Invited Plenary talk

Webometrics – Ten Years of Expansion

Peter Ingwersen

Department of Information Studies, Royal School of LIS, Denmark
pi@db.dk

The year 2006 commemorates the publication of the initial reports and journal articles on Webometric research, i.e., in the form of link analyses and distributions of websites over countries and domains. Springing from the broader Informetric and Scientometric research areas the attempts were to make measurements of the Web and somehow to try to define some useful and reliable indicators that inform us about something central of the information available on the Web.
The study of the Web was named 'webometrics' by Almind and Ingwersen (1997). It is defined as: "[The] study of the quantitative aspects of the construction and use of

information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches" (Björneborn & Ingwersen, 2004, p. 1217). One distinguishes between studies of the Web and studies of *all* Internet applications, commonly named 'cybermetrics'. Other alternative names have been suggested, such as, 'internetometrics' and 'webmetrics'. The latter conception has an increasing use on the web itself, see Table 1 below, but not in the scholarly communication among scientometricians. It seems also to be of broader meaning by including many analyses and measures other than associated with Informetrics. Table 1 demonstrates the distribution of the term 'webometric(s)' in SSCI and SCI in number of publications as well as the no. of publications that cites the original Almind & Ingwersen article (1997) from 1997. The right-hand side shows the distribution of the terms 'webometric(s)' and 'webmetric(s)' in one snapshot of four Web search engines.

Table 1. Distribution of the term 'webometric(s)' over articles in SCI/SSCI (1st column), the distribution of articles citing Almind & Ingwersen (1997) (2nd, column), and the snapshot of occurrences of webometric terms in four Web search engines (right). Analysis made on April 6, 2006.

		Citations to Almind	Google				
	"Webometric(s)"	*&Ingwersen, 1997*	Scholar	Yahoo	Google	Microsoft	Terms
2006	2	2	167	8.300	12.700	991	Webometric
2005	11	21	418	17.200	129.000	5.007	**Webometrics**
2004	14	18	17	565	62.700	640	Webmetric
2003	12	16	150	49.700	65.400	22.595	*Webmetrics*
2002	3	5					
2001	1	7					
2000	1	8					
1999	0	6					
1998	1	5					
1997	1	0					

In the citation indexes the term 'webmetric(s)' does not appear sufficiently often to merit a distribution analysis. We observe a certain stagnation of article volume in recent years but an increase in citations to the original article. To the right we observe that the term 'webometrics' is the most retrieved one in Google and Google Scholar, whereas in Yahoo and MSN.com the term 'webmetrics' is most powerful.

If we briefly look at the relationships between Informetrics, Scientometrics, Bibliometrics, Cybermetrics and Webometrics, Figure 1, we may observe that Webometrics associates with bibliometrics and overlaps Scientometrics to an extent (Björneborn & Ingwersen, 2004).

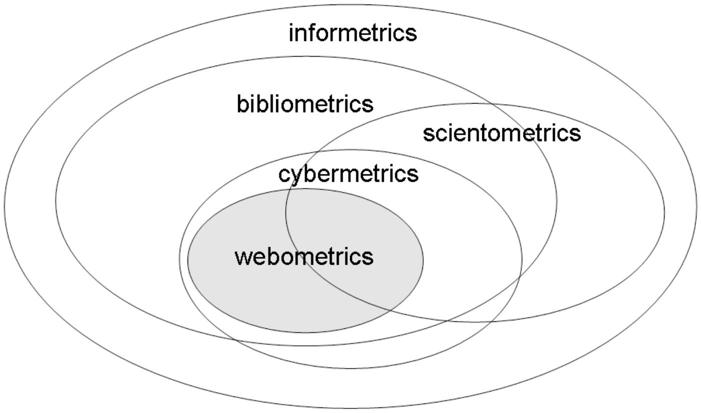

Figure 1. The relationships between the LIS fields of infor-/biblio-/sciento-/cyber-/webo-/metrics. Sizes of the overlapping ellipses are made for sake of clarity only (Björneborn, 2004).

For instance, simplistic counts and content analysis of Web pages can indeed be seen as analogous to traditional publication analysis; counts and analyses of outgoing links from web pages, named outlinks, and of links pointing to web pages, called inlinks, can be seen – perhaps quite erroneously – as somehow similar to reference and citation analyses. Since the open Web consists of contributions from anyone who wishes to contribute, its quality of information is often obscure. The Web most frequently demonstrates contents of non-scientific nature – se Table 2. Increasingly the margin between shading truth, misinformation, opinions, visions or speculation *and* reliability, validity, quality, relevance or truth becomes thinner. It becomes a matter of trust (Ingwersen, 2005).

Table 2. Quality assessment study of the Web (n: 3 x 200 = 600 sites). Application of several engines (Jepsen et al. 2004).

Category	**'plant hormones'**		**'photosynthesis'**		**'herbicide resistance'**	
	English	Scandinavian	English	Scandinavian	English	Scandinavian
Scientific	5%	1%	9%	0%	6%	1%
Scientifically related	17%	14%	25%	11%	24%	13%
Teaching	12%	37%	20%	19%	11%	16%
Low-grade	27%	12%	15%	16%	45%	48%
Noise	17%	15%	17%	41%	3%	1%
Unavailable	22%	21%	14%	13%	11%	21%

Having said that, the Webometric field has crossed the threshold of maturity by *not* ignoring these obstacles and difficulties. From the start of the new Century the field has expanded into several directions:

- *Web Indicators* – many laboratory groups around the world work to reinforce the quality of such measures and correlation studies;

- *Web space studies* – we observe a relationship to social networks and studies of other social phenomena of the Web;
- *Web data collection* – this area has since the start of Webometrics been of great importance, both in terms of quality assessments of search engines but also concerned with how to apply Web crawlers, adequate sampling methods, limitations as to webometric analyses, etc. There is here, as in other Informetric analysis work, a strong link to the field of information retrieval (IR) – here in terms of Web IR and Web data mining;
- *Web link analyses* – in particular in (well) defined (scientific) domains.

The correlation analyses are quite interesting since we take established S&T indicators and observe how novel Webometric indicators may provide insights. We are aware that a central difference from traditional scientific databases and archives is the dynamics of the Web. *Time* plays a different role on the Web.

Further, as stated in (Ingwersen, 2005) and by other Webometric researchers: links are *not* necessarily *normative*, such as credit granting or recognition providing devices, but *functional* in nature. We do not know for sure why people on the Web link up to other pages – but there are many reasons detected in our analyses (Wilkinson et al., 2003). There exist no conventions of linking as in the scientific world. Negative links are almost non-existing. The analogy between links and references or citations is hence of the superficial kind and should definitively not be taken too far. For instance, there seem to exist statistically significant correlations between number of inlinks and productivity at academic Web sites – not between inlinks or Web Impact Factor and peer reviewed quality of the production – as might be expected at first (Thelwall, Vaughan & Björneborn, 2005). Nevertheless, the same analogy may indeed provide interesting hypotheses about the characteristics of links and their meaning – as originally done by Brin & Page (1998) with the Google PageRank algorithm. Here, inlinks are seen as providing cognitive authority.

In sum, we can observe that the number of articles on Webometric issues has increased dramatically, not simply publications containing the term. We have now got published monographs in the field as well as textbooks and larger review articles in central review journals. At least one professor has been appointed in the field. We have demonstrated many efforts in order to ensure validity and significance of Webometric studies.

The next step is to face how to merge or federate parts of the open Web, dedicated scientific Web segments, like Google Scholar and CiteSeer and/or open as well as closed repositories and archives belonging to publishers and universities for research evaluation purposes with mixed indicators. This requires novel or alternative data collection methods and fusion techniques for performing
reliable and sophisticated analyses.