Web mediators for cybermetric purposes: A comparative analysis

Isidro F. Aguillo, José Luis Ortega, Mario Fernández, Ana M. Utrilla
Cybermetrics Lab. CCHS – CSIC
Joaquín Costa, 22. 28002 Madrid. Spain
{isidro;jortega,mariofdez,ana.utrilla}@cindoc.csic.es

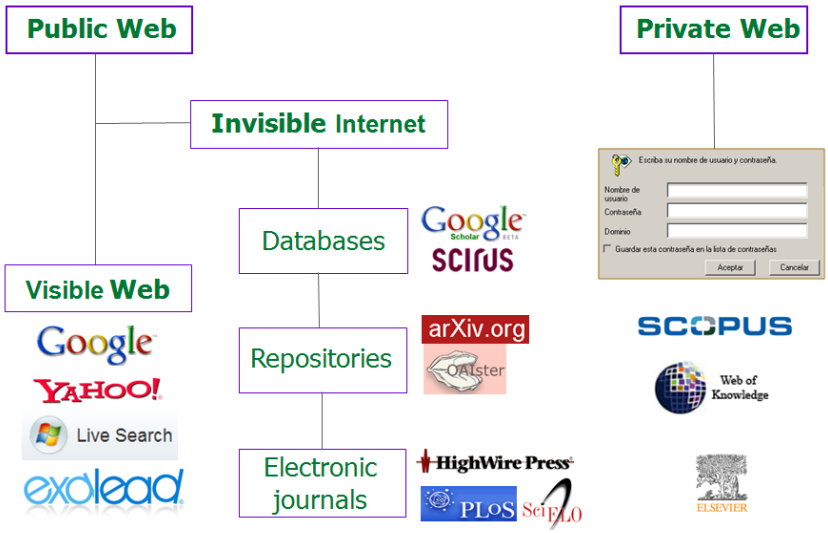
2008-7-2

Webometrics is becoming one of most relevant informetric sub-discipline (Bar-Ilan, 2008) with an important activity in several fields complementing traditional bibliometric approaches or starting new developments. The availability of trusted web data is of paramount importance to achieve good results but there are technical problems related to the tools or intermediaries required for web data recovery.

The aim of this communication is to show a classification and empirical analysis of the main Web mediators according to their cybermetric capabilities. Summarizing the main advantages and shortcomings of these intermediaries allow to suggest best practices in the design and execution of future webometrics studies.

Web mediators.

For practical reasons we will focus on two complementary areas of the webspace, the visible web defined as the electronic information available from the databases of the commercial search engines and the invisible web, usually including the data not collected by the search engines robots due to structural reasons, although here restricted to those databases in which the access gateway imposes a barrier to these robots.



The first group has evident mediators as the commercial search engines clearly define its limits. Obviously, specially designed crawlers could be used (Thelwall et al.) to recover additional information from a website, but they are difficult to customize and they do not offer significant advantages (Cothey et al.) over commercial search engines. Moreover, engines are ubiquitous and universally used in web sessions for accessing academic information.

The number of large search engines with independent databases is notoriously small. In fact there are only 8 of them, and even fewer have strong cybermetric capabilities. In this analysis we will focus on Google, Yahoo Search (and mirrors), Live Search and Exalead.

The invisible Web is no longer a homogeneous group as several search engines, specially Google are being able to index a large part of its contents. For the purposes of our analysis we recognize three large components: Academic databases (Academic Live, Google Scholar, Scirus, ..); Repositories, including meta-services (aggregators, directories) and Electronic Journals. As the study of electronic journals is basically bibliometric, this group has been excluded.

Methodology.

For the empirical analysis two sets of academic web domains have been used. The catalogue of Universities of the Webometrics Ranking (www.webometrics.info) consists of 14877 institutions of higher education worldwide with autonomous web domains. From the same source, a list of academic repositories has been extracted, including 590 thematic and institutional repositories from all over the world.

For each domain, quantitative data has been extracted from both general and academic search engines. From commercial search engines the number of web pages and rich files (pdf, doc, ps, ppt and xls) are obtained, whereas in Google Scholar the number of items (papers in several formats and citations) are collected. For comparison purposes, two consecutive rounds were used for each engine and topic.

Results.

Several authors have noted the inconsistencies of the search engines. Our results confirm that there are important differences in the two rounds used for most of the search engines, except Google Scholar that it is very stable.

However most of these differences are due to “rounding” procedures: Google only offer figures ending in ’0, ‘00 or ‘000, while Yahoo adjust the number of results better in the last pages of showed results. The data also confirms that not all the Google datacenters provides the same values. Furthermore, the databases accessed by search engines API’s are smaller than the commercially versions. As expected the comparative analysis of the Yahoo mirrors show that both AltaVista and Alltheweb are using the same database and providing similar results (Chi-square test, $p=0.000$) allowing the use of these alternative sources for large projects.

Regarding geographical coverage, there are significant differences between the search engines. Exalead, a French engine, show a strong European bias whereas Live, currently the largest one, show a North American preference. Asian coverage by Google is far better than the others. Comparing Scholar with the main database there is a striking low coverage of Asian academic space, but it is more consistent with the data provided by other search engines. A similar pattern arises when rich files are used in the comparison.

REGION	SIZE				SCHOLAR	RICH FILES			
	GOOGLE	YAHOO	LIVE	EXALEAD		GOOGLE	YAHOO	LIVE	EXALEAD
NORTH AMERICA	34.8%	47.7%	57.4%	49.5%	38.6%	41.8%	47.8%	64.2%	53.6%
EUROPE	34.5%	31.9%	28.1%	42.0%	39.8%	36.3%	35.4%	23.0%	37.3%
ASIA	23.6%	14.9%	8.7%	4.5%	10.1%	12.5%	10.4%	6.5%	3.6%
LATIN AMERICA	5.1%	3.0%	2.5%	2.0%	8.4%	6.1%	3.8%	3.1%	2.5%
OCEANIA	1.3%	2.1%	2.8%	1.6%	2.5%	2.0%	1.8%	2.5%	2.3%
AFRICA	0.4%	0.3%	0.3%	0.3%	0.4%	0.6%	0.3%	0.5%	0.3%
MIDDLE EAST	0.3%	0.2%	0.2%	0.1%	0.3%	0.7%	0.3%	0.2%	0.3%

COUNTRY	SIZE				SCHOLAR	RICH FILES			
	GOOGLE	YAHOO	LIVE	EXALEAD		GOOGLE	YAHOO	LIVE	EXALEAD
USA	15.7%	22.2%	26.5%	45.1%	17.7%	38.5%	44.4%	61.4%	48.6%
GERMANY	3.0%	3.3%	2.9%	8.5%	3.0%	7.5%	6.6%	5.1%	8.3%
UNITED KINGDOM	1.5%	1.9%	2.9%	6.2%	1.8%	3.5%	3.5%	3.9%	4.5%
FRANCE	0.8%	0.8%	1.0%	7.3%	0.9%	1.8%	1.7%	1.5%	6.3%
CANADA	1.7%	1.6%	2.2%	4.4%	1.6%	3.3%	3.5%	2.8%	5.0%
CHINA	3.2%	2.7%	1.0%	1.4%	0.4%	2.5%	1.8%	0.8%	0.4%
JAPAN	2.1%	1.8%	1.6%	1.4%	1.8%	3.7%	3.3%	3.0%	1.5%
SPAIN	1.2%	1.4%	0.8%	2.0%	4.3%	3.6%	2.4%	1.6%	2.6%
ITALY	1.0%	0.7%	0.8%	2.5%	1.1%	2.8%	2.4%	2.1%	2.8%
AUSTRALIA	0.6%	1.0%	1.2%	1.4%	1.1%	1.7%	1.6%	1.9%	2.0%

The volume of data available from the repositories is still very small. Only the large disciplinary repositories (Arxiv, CiteSeer) using extensively Postscript (ps) format are well represented. As it could be expected Scholar is including already these contents in its database.

SEARCH ENGINES	SIZE	PDF	DOC	PPT	PS	XLS	SCHOLAR
GOOGLE	5,21%	4,31%	0,19%	0,24%	0,80%	0,09%	31,54%
YAHOO	3,84%	2,76%	0,11%	0,11%	27,62%	0,05%	
LIVE	1,82%	1,92%	0,19%	0,26%	n.a.	0,12%	
EXALEAD	2,00%	0,46%	0,11%	0,12%	n.a.	0,03%	

Conclusions

The main recommendation is that as none of the search engines is the best for cybermetric purposes since they offer different capabilities and biases, webometrics analyses should use all of them to compensate or at least decrease the problems shown.

Google Scholar is very stable and due to its citation capabilities offers new possibilities for bibliometric/cybermetric analysis.

The repositories are not still a comprehensive source of data for global analysis.