



WebTAPIR – Scientific Information Retrieval on the World Wide Web

Erik Thorlund Jepsen, Piet Seiden, Lennart

Bjørneborn, Haakon Lund & Peter Ingwersen

{etj;ps;lb;hl;pi}@db.dk Department of Information Studies, Royal School of Library and Information Science, Denmark

Aim

WebTAPIR (Web based Text Access Potentials for interactive Information Retrieval) is a project aimed at developing filtering and ranking methods in order to enable users to discriminate between web resources containing scientifically relevant content versus other content.

Research goals

- » Establishment of a scientific web page test collection.
- » Development of filtering algorithms focusing on the utilization of poly-representations and 'cognitive retrieval overlaps' (Ingwersen, 1992 & 1996).
- » Testing of developed filters in relation to interactive information retrieval.

Search results

Three search engines were employed as a means of generating a preliminary test collection. For each subject and language grouping, multiple searches were performed in order to retrieve synonyms as well as both singular and plural forms of the terms. The search engines employ restrictions on the number of hits one is allowed to retrieve. Consequently, these cut-offs result in only the Nordic language group getting close to being exhaustively searched here.

Subject	Google	AllTheWeb	AltaVista
Cut-off	appr. 1,000	4,100	200
Photosynthesis			
Nordic	6,523	3,837	3,995
English	280,500	146,270	104,571
ratio	2.3%	2.6%	3.8%
Herbicide resistance			
Nordic	360	341	292
English	73,318	40,324	33,668
ratio	0.5%	0.8%	0.9%
Plant hormones			
Nordic	244	358	159
English	299,900	17,640	134,299
ratio	0.1%	2.0%	0.1%

The subjects chosen were expected to occur in widely differing treatments.

- » Photosynthesis: Teaching materials, high school reports, popular science, scientific content.
- » Herbicide resistance: Technical reports, NGO-sites, scientific materials.
- » Plant hormones: Scientific content, agronomical use of growth regulating compounds.

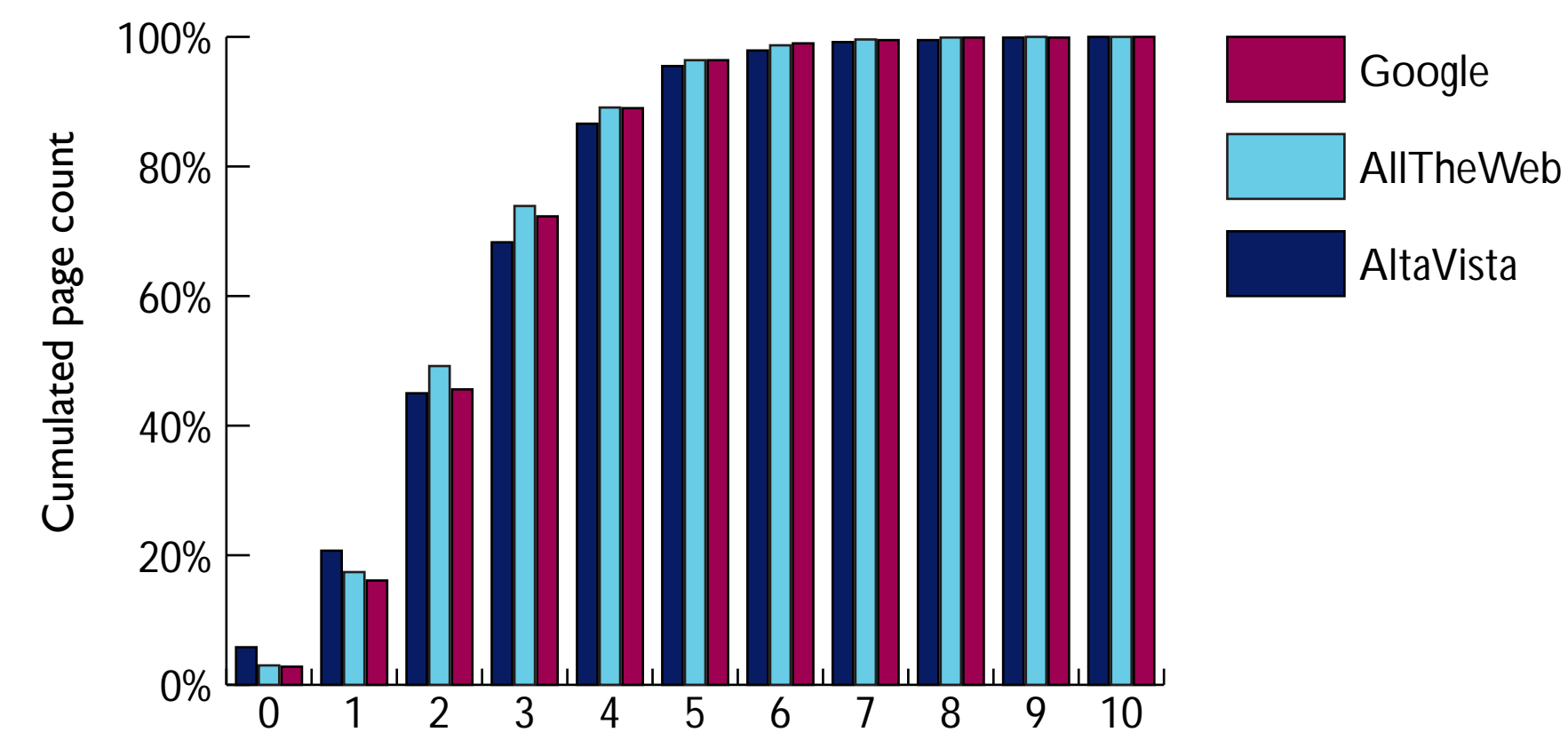
URL's retrieved

Photosynthesis URL's	Google	Alltheweb	Altavista	total	% of total
Nordic	2238	3273	630	6141	37%
English	1712	8200	400	10312	63%
total	3950	11473	1030	16453	100%
Language overlap	1	12	0		
Search Engine overlap					
with Google	x	1326	351		
with Alltheweb	1326	x	518		
with Altavista	351	518	x		
with (Alltheweb+Altavista)	1404	x	x		
with (Alltheweb+Google)	x	x	596		
with (Google+Altavista)	x	1466	x		
overlapping documents				3466	21%
% unique URL's	64%	87%	42%		

The retrieved URLs showed some overlap between the chosen search engines, based on analysis of the URLs, but much less than expected, with a worst case of only 42% unique URLs, as seen for AltaVista in the photosynthesis dataset shown above.

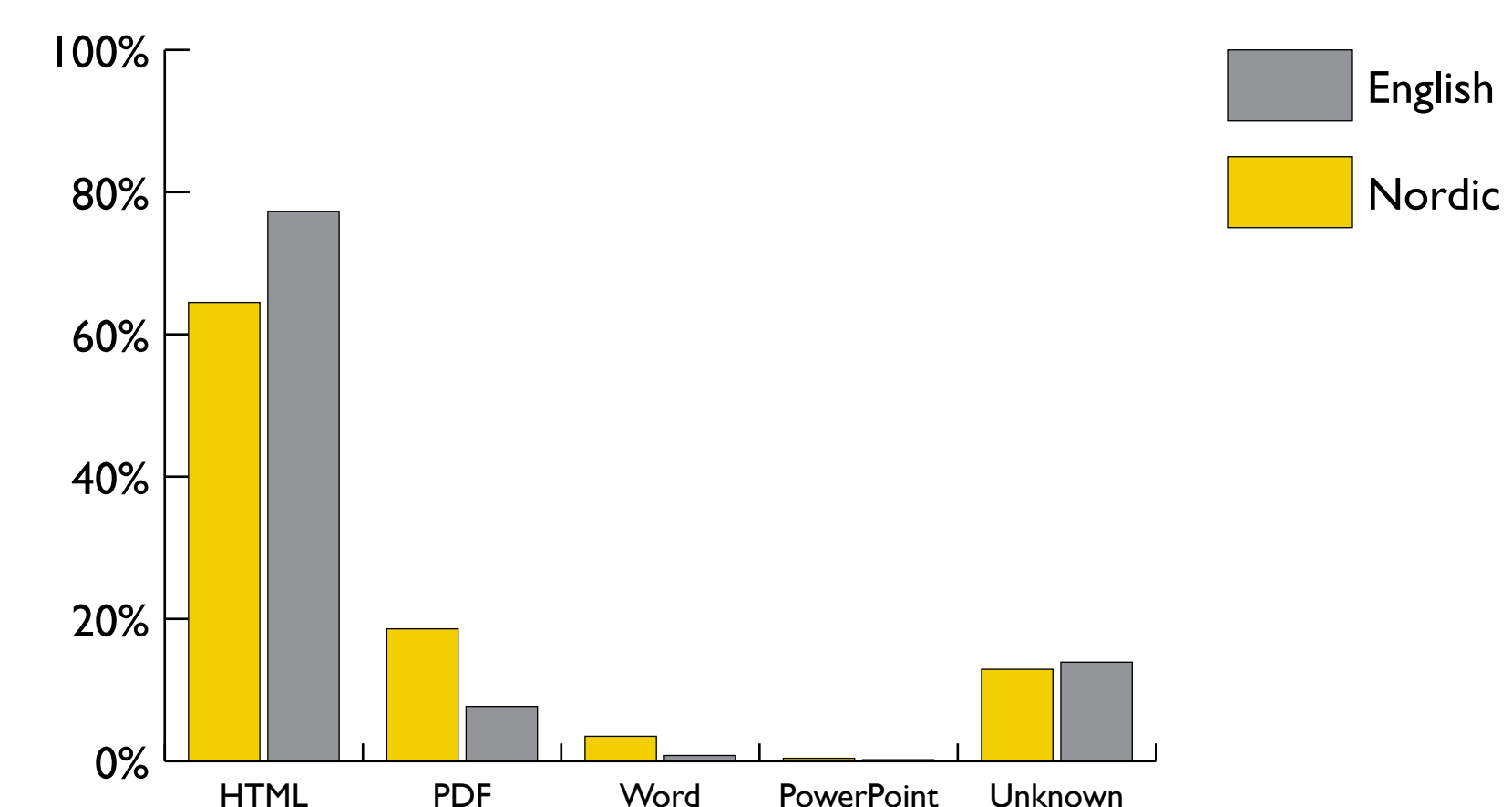
This confirms that the search engines index different sites (Bar-Ilan, 1999). But they may also differ quite substantially in their methodology used for ranking of the results.

Search engine indexing levels



The search engines appear to behave similarly with regard to the level of webserver file systems that is indexed, giving a similar distribution of URLs when categorized on basis of the URL-structure with regard to file system traversal. As can be seen from the figure above there is no discernible difference between the search engines.

Document formats



Based on the results obtained from Google, the prevalence of varying document types could be established, as Google indexes a range of popular formats. Apart from HTML only PDF documents are also significant in their occurrence, especially for the Nordic language group. This could be taken as an indication that PDF documents are highly relevant for scientific subjects, as has been suggested by Lawrence and Giles (1999).

Conclusions

These preliminary investigations have provided a starting point for the development of filtering methods. The search engines utilized only returned the highest ranking fraction of the found URLs and could thus only be used as tools for generating a skewed sample of the documents expected to reside on web servers. But by analyzing a subset of the retrieved URLs, a basis for preparing filtering methods has been established, which then can be further tested as regards improving relevance in information retrieval of scientific information on the World Wide Web.

References

- Allen, Eva S., Burke, John M., Welch, Mark E. & Rieseberg, Loren H. (1999). How reliable is science information on the Web? Nature, 402, p. 722.
- Bar-Ilan, J. (1999) Data collection methods on the Web for informetric purposes - A review and analysis. Scientometrics 50 (1) p. 7-32.
- Bjørneborn, Lennart & Ingwersen, Peter (2001) Perspectives of webometrics. Scientometrics, 50 (1), p. 65-82.
- Cronin, B. & McKim, G. (1996). Science and scholarship on the World Wide Web: A North American perspective. Journal of Documentation, 52, 163-172.
- Ingwersen, Peter (1992). Information retrieval interaction. London: Taylor Graham.
- Ingwersen, Peter (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive IR theory. Journal of Documentation, 52 (1), 3-50.
- Lawrence, S. & Giles, C. L. (1999). Accessibility and distribution of information on the Web. Nature, 400, 107-110.
- Sullivan, Danny (ed.). (2001). Search Engine Features for Searchers. <http://searchenginewatch.com/facts/ata glance.html> (Page updated: Oct. 26, 2001)