Articles, Papers, Chapters, Theses – who wins the Visibility Wars?

Melius Weideman
Website Attributes Research Centre
Cape Peninsula University of Technology
P.O. Box 652, Cape Town, South Africa
27 21 9135515
weidemanm@cput.ac.za

ABSTRACT
Researchers need access to previous research to base their own work on. Some of the most commonly referenced materials are published in the form of journal articles, conference papers, books and book chapters, and research theses. The purpose of this research was to determine how these four categories of documents compare in terms of visibility to search engine crawlers. A questionnaire was used to gather data from international scholars on their completed research. Three types of queries were generated and over 3000 websites were inspected to determine the visibility of these outputs. Search engine result pages were inspected, and the rankings of the research documents were recorded and converted to a scoring system. The results have indicated that the four types of outputs enjoy varying degrees of exposure to search engines, with journal articles leading the way, and books/book chapters having the smallest degree of exposure to search engines. Some query types also produced better results than others. It was concluded that journal articles provide the best way to expose research work to Internet searchers through search engines.

Keywords
journal article, conference paper, book, thesis, search engine, ranking

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Query formulation, Information filtering

General Terms
Algorithms, Measurement, Experimentation.

1. INTRODUCTION
Some modern research is based on work done by other authors in the past. For this to be possible, these other publications must be accessible – preferably the full-text and not just the abstract. A large part of this body of knowledge is accessible through a myriad of academic databases, most of them through a subscription service. As a result, many researchers attempt to rather seek these publications through a free service, especially if their institutional library does not subscribe to a specific academic database. Free-from searching in this research is defined as being the use of search engines to find academic data, without registration or payment.

This trend has been called "Open Access", referring to the fact that any user can access the full content of published research work without paying for the privilege [4].

The use of search engines to find answers to questions, with a consistently high success rate, remains to be an evading target [5]. Furthermore, digital libraries, while offering easily accessible academic content, rely on the premise that users will access this content via the home page and drilling down through menus. However, many users start a search through a search engine, resulting in access to content via non-homepages.

The purpose of this research was to determine to what degree completed research publications are accessible via free-form searching, and to determine which type was more visible.

2. OTHER RESEARCH

2.1 Search engine ranking
Humans have always been involved in ranking objects and other entities, with one being considered "better" than another - whether done intuitively or based on actual measurement. Any ranking is automatically subject to scrutiny, especially by those who do not agree with its result. A typical example is the continuous ranking of universities, whether it is on a by-country or global basis. Table 1 lists some of these ranking structures.

2.2 Free-form Searching
Seiler *et al* confirmed that users of a website selling property first looked at the photographs before reading the descriptive text [3]. However, a finding that corresponds to search engine result page readers' tendency to sometimes read only top results, was that prospective home-owners spent significantly more time on viewing the first house on a webpage than those lower down.

Another recent study focused on the problem of finding specific expertise on the social Web [1]. One finding of implementing a new service of this type was that users would rather get an answer quickly, even when slightly incomplete, than wait longer for a more detailed answer. This confirms earlier research that users are generally impatient, and also the importance of ranking highly for a given key-phrase.

3. METHODOLOGY
A number of issues had to be addressed before executing this research. These include the form of data collection, search engines to be used, and the generation of effective search queries.

978-1-4799-5569-5/14/$31.00 ©2014 IEEE

Table 1: Three global university ranking schemes.

Name	Ranking URL
ARWU	http://www.shanghairanking.com/
QS Top Universities	http://www.topuniversities.com/
Webometrics	http://www.webometrics.info/

3.1 Data Collection

In an attempt to gather enough metadata about research publications, a questionnaire was designed and sent to a number of international authors. Data was requested on four types of publications: journal articles, conference papers, books or book chapters and student theses (masters and doctoral).

The following metadata was requested: Title, Year of publication, Author surnames, Abstract text, First paragraph body text and Keywords.

3.2 Search Engines

The latest statistics in core search volume indicate that the trend of the previous year is continuing: Google leading with about a 2/3 market share, but Bing in a distant but growing second place, and Yahoo! falling behind constantly into third place [2]. It was clear that if these three were used in this research, an acceptably large section of the market would be included.

3.3 Search Queries

It was decided to use three search queries, combining the best practice results of previous studies. Q1 would be the first sentence of the abstract and Q2 the first sentence of the body text. The third query would be a concatenation of the publication title, the author surnames, year of publication and keywords.

The number of searches to be executed to obtain data for this research was: (3 Search Engines) x (3 Queries) x (42 unique documents) = 378. For each search, the first 10 results were inspected to determine whether or not the full-text was accessible after the first click. A total of 3780 websites were therefore visited and inspected during the empirical research. A score of 10 was allocated for first position, 9 for second, down to 1 for 10th position.

4. RESULTS AND ANALYSIS

The 18 valid data sets (out of 25 respondents) contributed a total of 42 valid research publications. Authors were identified by unique numbers and publications by single letters (j, c, b and t for the four types respectively) – see Table 2. One example of each publication type is also shown for clarity. For example, document 22c column GSQ1 refers to refers to Google search query no. 1 producing a ranking of 1 (= score of 10).

Table 2: Extract of Results

No. a	No. b	GSQ1	BSQ1	YSQ1	GSQ2	BSQ2	YSQ2	GSQ3	BSQ3	YSQ3	
21	b	0	0	0	0	0	0	0	0	0	
		1.3	1.1	1.3	1.6	2.3	2.4	2	2.1	2	1.776
22	c	10	10	0	10	10	0	10	10	0	
		5.7	3.5	1.3	4.5	1.8	0	5.6	5.6	6.3	3.794
16	j	9	8	7	9	8	7	10	9	8	
		6.2	5.7	4.4	7.3	5.2	5.6	7.9	6.2	6.8	6.138
17	t	9	0	0	9	0	0	10	10	10	
		1.8	0	0	3.8	-0	-0	7.2	2	2	1.822

5. CONCLUSION

It is clear from the results that journal articles have the highest degree of visibility to search engines, with a score of 6.138. They are followed by conference papers with 3.794, then theses with 1.822 and the least visible were books and chapters with 1.776. The reason why the top two have a high score is assumed to be the increase in open access to research publications, which is predominant in the academic world for these two types of publications. Books, being last, are assumed to have a low degree of visibility due to copyright and high price issues. Academic researchers requiring high exposure on the open Internet are advised to concentrate on journal articles and conference papers for their research publications.

6. REFERENCES

[1] E.H. Chi, "Who Knows? Searching for Expertise on the Social Web," Communications of the ACM, Research Insights, vol. 55 no. 4, pp. 110, 2012.

[2] A. Lella. (2014, Jan.). comScore Releases January 2014, U.S. Search Engine Rankings. [Online]. Available: http://www.comscore.com/Insights/Press_Releases/2014/2/comScore_Releases_January_2014_US_Search_Engine_Rankings

[3] M.J. Seiler, P. Madhavan and M. Liechty, "Toward an understanding of Real Estate homebyuer Internet searching behaviour: An application of ocular tracking technology," Journal of Real Estate Research, vol. 34 no. 2, pp. 211 – 241, 2012.

[4] P. Suber, (2013). Open Access Overview. [Online]. Available: http://bit.ly/oa-overview

[5] M. Weideman, Website Visibility: The Theory and Practice of Improving Rankings. Oxford: Chandos, 2009.