Benchmarking Google Scholar with the New Zealand PBRF research assessment exercise

Alastair G Smith
School of Information Management
Victoria University of Wellington
New Zealand
alastair.smith@vuw.ac.nz

Abstract
Google Scholar was used to generate citation counts to the web-based research output of New Zealand Universities. Total citations and hits from Google Scholar correlated with the research output as measured by the official New Zealand Performance-Based Research Fund (PBRF) exercise. The article discusses the use of Google Scholar as a cybermetric tool and methodology issues in obtaining citation counts for institutions. Google Scholar is compared with other tools that provide web citation data: Web of Science, SCOPUS, and the Wolverhampton Cybermetric Crawler.

Introduction
An issue in cybermetrics is the extent to which evaluation measures for institutions based on web citations correlate with measures derived from a formal reseach assessment exercise (for example (Thelwall & Harries, 2003), (Smith & Thelwall, 2002)). This study uses Google Scholar (http://scholar.google.com/) to compare the web citation rate of New Zealand universities with the results of New Zealand's Performance Based Research Funding (PBRF) research assessment exercise, and discusses the utility of Google Scholar as a cybermetric tool.

A number of researchers (for example (Thelwall, 2002), (Bar-Ilan, 2005)) have pointed out that general web link counts from search engines are misleading, due to the lack of transparency in the algorithms used to arrive at search counts, the inability to reproduce results as search engines make arbitrary changes to their algorithms, etc. Also, general link counts include much material which occurs on university websites but is unrelated to the university's research objectives (Thelwall, 2003b). Examples include teaching materials, general administrative materials, and personal websites of staff and students (for example, at one time much of the web traffic to the author's institution was to a recipes database maintained as a hobby by a chemistry PhD student).

Google Scholar has been introduced as a research oriented web search engine. The web sites crawled are selected research oriented sites, for example electronic journals (both open access and subscription based), online theses collections, research reports, and preprints. Google Scholar not only provides full text searching of this material, but extracts formal citations from the material. This means that Google Scholar acts as a citation index as well as a search engine. Also, Google Scholar extracts citations to print materials that have been referenced in web publications. This means that the database provides access to some material in the conventional print environment, as well as to web based material.

There are criticisms that neither the sources, nor the selection criteria, are made public (e.g. (Jacsó, 2005)), but nonetheless Google Scholar has become widely used for searching research oriented material. An evaluation of a number of searches for research material in the academic environment concluded that Google Scholar was a useful reference tool for academic librarians (White, 2006). A study of Google Scholar showed that it contained citations to significant resources that were not covered by the Science Citation Index database (Kousha & Thelwall, 2006).

It has been suggested (Noruzi, 2005) that Google Scholar has potential as a citation index for bibliometric work. Citation counts for JASIST articles were found to be higher in Google Scholar than in ISI's Web of Science or Scopus (Bauer & Bakkalbasi, 2005). Google Scholar has been compared with Web of Science and Scopus for calculating the h-index for highly cited Israeli researchers (J. Bar-Ilan & Lin, 2006). A comparison of citations to LIS literature using Google, Google Scholar, and the ISI's Web of Science indicated that Google Scholar had potential to replace ISI as a source of science and technology indicators (Vaughan & Shaw, 2006), although the current implementation of Google Scholar was inconsistent.

In 2003/4, New Zealand's Tertiary Education Commission (TEC) undertook a research assessment exercise (Tertiary Education Commission, 2004), evaluating the research output of New Zealand tertiary institutions, including the eight New Zealand universities. Academic staff at the institutions submitted portfolios in which they reported their research outputs for the previous six years (which included publications, contributions to the research environment (e.g. editing journals, organising conferences), and indications of peer esteem (e.g. awards). Subject based panels awarded grades to portfolios, which were then translated to a numeric score that indicated the output of the staff member. Quality scores (an average output per staff member) and total outputs have been published for each institution. There have been some studies which criticise the effectiveness of PBRF and its impact on teaching and research at universities (for example (Morris Matthews & Hall, 2006)) but the PBRF ratings provide a benchmark with which to compare bibliometric or cybermetric measures.

In a previous study, PBRF rankings were compared with inter-university links obtained by a specialised crawler (Smith & Thelwall, 2005). This study found a moderate correlation between link counts per FTE staff member and the PBRF quality score, reinforcing the idea that a cybermetric study could produce a measure similar to an established research evaluation measure. An issue with the Smith and Thelwall study is that the counts were limited to links between the NZ universities, so did not reflect international linkages. International linkages are significant in a relatively small country such as New Zealand. Also, the link counts included non-research material, since the crawler covered all material in the universities' domain that it could reach, without regard to whether the material was research oriented.

Methodology:

The current study uses Google Scholar to obtain web citation counts to research originating from the eight New Zealand Universities. Google Scholar includes research related material from the Web, along with print items cited in Web documents. A web citation count is provided for each item displayed in a search. In the current study Google Scholar was searched to provide a set of research associated

with each university, and a citation count extracted. This citation count was then compared with the total PBRF output for the universities.

Two methodology problems occurred with Google Scholar:

- Identifying work related to a university. Unlike Web of Science, for example, Google Scholar does not have an institutional name field.
- Obtaining a total citation count. Google Scholar only displays a citation count for each item, and only allows the first 1000 hits to be displayed from a search.

The most effective search for a university was on the name of the institution, combined with city/province names where required to remove ambiguity. So for example a search for material originating at Canterbury University was:

```
"canterbury university" OR "university of
canterbury" zealand OR christchurch OR ilam
```

since Canterbury University is located in the suburb of Ilam in the city of Christchurch. This formulation removed hits on, for example, "University of Kent at Canterbury".

A search on domain name (e.g. `host:www.vuw.ac.nz`) seems obvious, but produced many false drops (e.g. research from other institutions mirrored at the university) and missed research that was not hosted at the originating university (e.g. papers presented at external conferences).

Google Scholar does not provide a direct method of determining the total citation count for a set of items. It only provides a citation count in the entry for each item. In the study a macro was used to extract and total the citation counts from the Google Scholar result lists. An additional issue is that Google Scholar only allows the first 1000 items from a search to be shown. However citation counts appear to be used in the ranking algorithm, and items ranked near 1000 tend to have few or no citations, indicating that most of the institution's citations will be in the first 1000 hits. The study assumed that the citation count from the first 1000 items is a reasonable approximation of the total citation count for the institution.

Results:

Table 1 shows for each institution: the total hits (number of items reported by Google Scholar for each university), the number of citations to the first 1000 items retrieved by Google Scholar, and the total PBRF output (the PBRF quality score multiplied by the number of full-time equivalent staff).

Correlation measures (Excel CORREL function) were used to compare PBRF and Google Scholar based measures. The best correlation appears to be between the total PBRF output and the total citations (0.94). This is also illustrated in Graph 1. This gives some support to the idea that citations as measured by Google Scholar are a useful measure of the research output of an institution, despite the methodological problems associated with the search engine.

There is also a reasonable correlation between the total PBRF output and the total hits reported by Google Scholar (0.85). This is illustrated in Graph 2. An interesting aspect of this graph is that the outlier corresponds to Otago University. Otago has

research strengths in medicine and biosciences, where research publication is likely to be in conventional print media rather than in the web sources covered by Google Scholar. This could explain the result that Otago's PBRF research outputs are comparatively higher than the web based outputs counted by Google Scholar.

Of course, Google Scholar is not the only tool available for cybermetric analysis of research publication on the Web. Web of Knowledge (the web interface to ISI's citation databases), and SCOPUS (Elsevier's recently introduced citation database), both index research based web publications to some extent. The University of Wolverhampton's dedicated cybermetric crawler (Thelwall, 2003a) also indexes links made between university websites.

As part of the current research, a comparison was made between these different tools for cybermetric research, and the results summarised in Table 2.

A key point is that Google Scholar accesses more material than the Web of Knowledge, SCOPUS, or the cybermetric crawler. An advantage of Web of Knowledge is that the field structure is more finely grained, for example there is an institution name field that enables output from a specific university to be identified. The cybermetric crawler allows identification by domain name, but as noted above, this may not necessarily indicate research outputs of the institution. While there are limitations to the ability to generate citation counts from Google Scholar, a casual user faces difficulties in obtaining citation counts from Web of Knowledge and SCOPUS – for example SCOPUS limits citation analysis to a relatively small number of hits, which makes citation analysis feasible for individual authors, but not for institutions. At the time of writing, SCOPUS had just introduced a "webcites" feature which may be useful for cybermetric analysis of author's work. A concern in using Google Scholar for cybermetric work, however, must be the lack of transparency in the search algorithms used, and the selection of material for crawling.

Another aspect arising from this study is the increasing availability of institutions' research output through institutional repositories. If these repositories are structured in such a way as to enable crawling by, for example, Google Scholar or the Wolverhampton cybermetric crawler, these repositories may become an important source of cybermetric data.

Conclusion

Google Scholar has shortcomings as a cybermetric tool, for example lack of transparency in algorithms and scope, and lack of cybermetric oriented search functions, such as an institutional name field and a method of obtaining true overall citaion counts. However this study indicates that Google Scholar provides a good coverage of research based material on the Web, and a relatively simple method of deriving measures of research output, for example citation counts for web based material. Google Scholar's measures correlate with a conventional research assessment exercise, the PBRF. This means that Google Scholar provides a relatively simple way of assessing the Web based research output of institutions. As more research information is available on the Web, for example through the development of institutional repositories, Google Scholar may be a significant tool for cybermetric work.

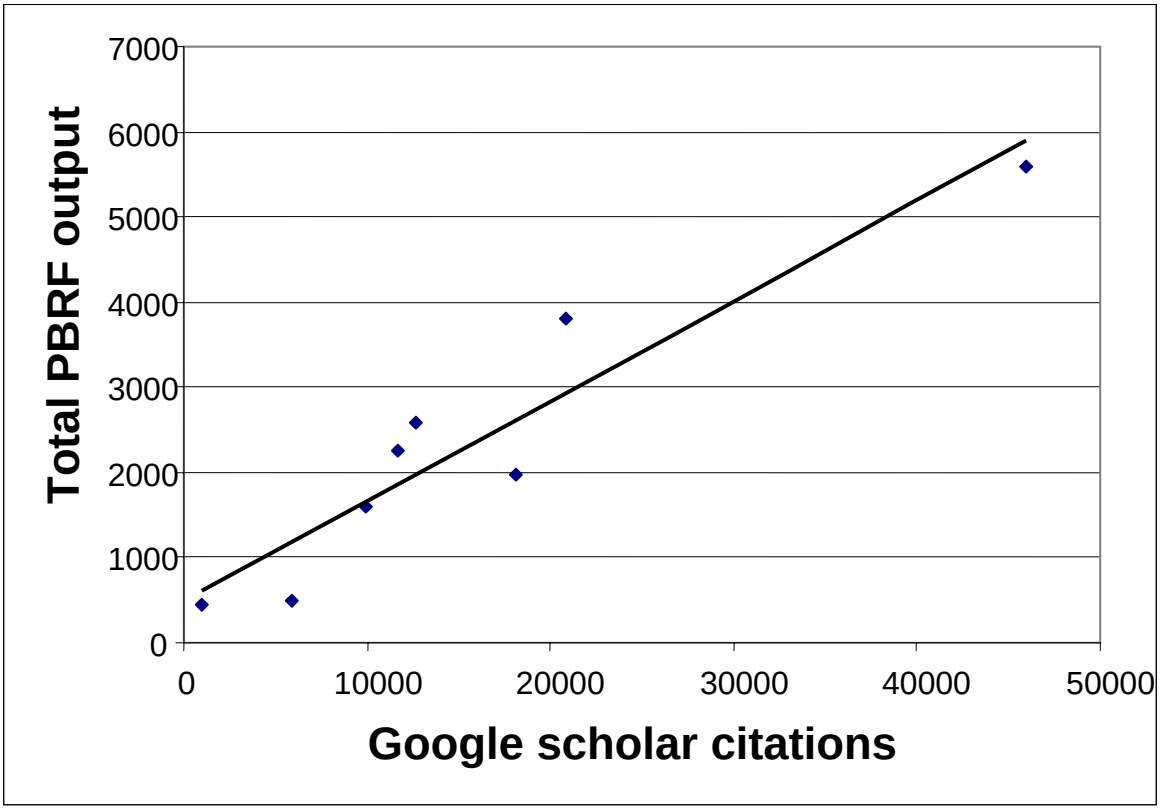
Table 1: Google Scholar and PBRF indicators for NZ Universities

Institution	Auckland	Auckland University of Technology	Canterbury	Lincoln	Massey	Otago	VUW	Waikato
Total hits	51500	1490	12300	4630	16100	13000	23100	15700
Total citations	45956	1028	11716	5920	12670	20890	18068	9985
PBRF output	5591	437	2260	500	2586	3795	1964	1598

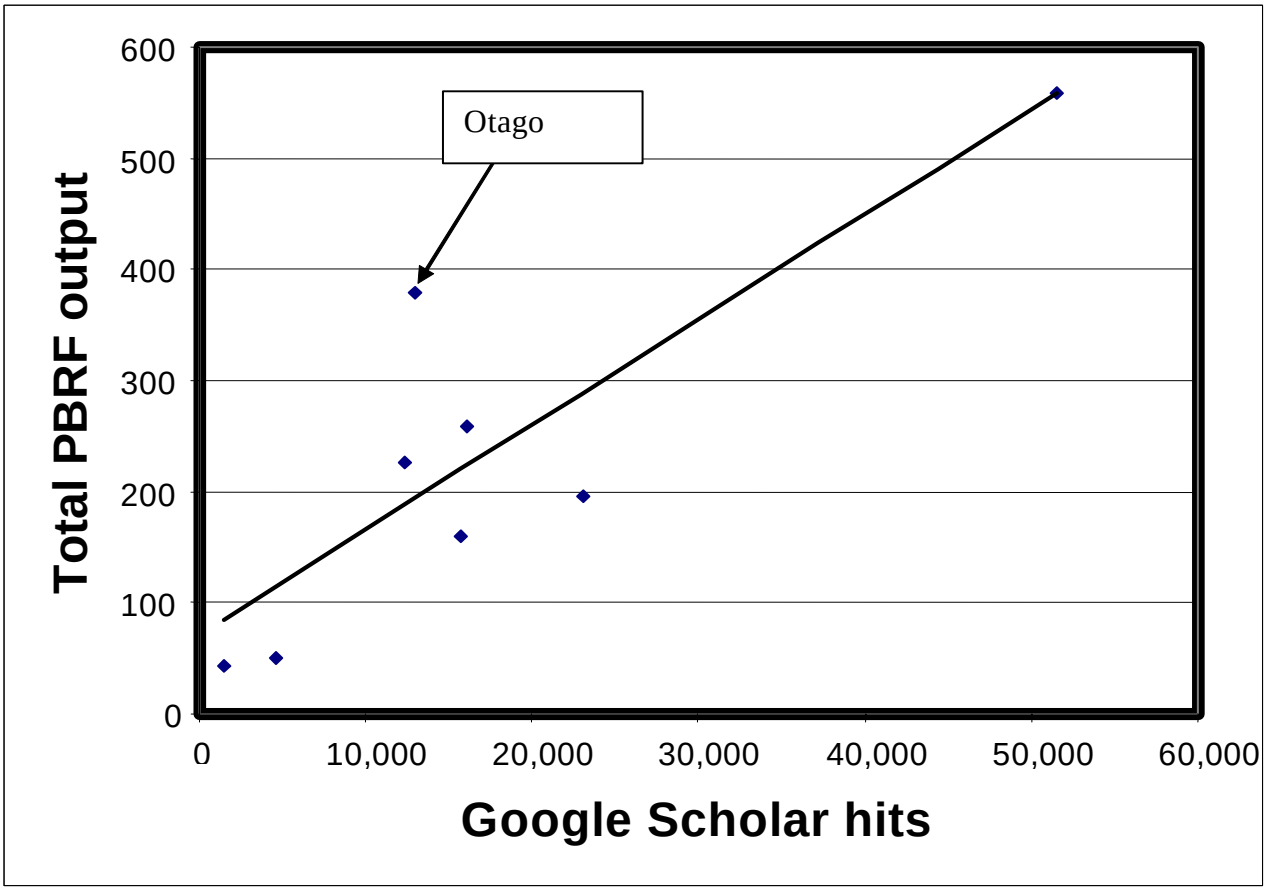
Table 2: Comparison of Google Scholar with other citation tools

	Coverage	Identification of institutions	Citation count	Transparency
Google Scholar	Research on Web	By keyword	Individual, cannot display all hits	Little documentation of algorithm, selection of sources
Web of Knowledge	Core journals (some digital)	Specific field	For individual items	Sources documented
Scopus	Core journals + Web sites (from Scirus)	Specific field	"Citation tracker" only for limited number of hits	Sources documented
Wolverhampton Crawler	Specific university web sites	By domain	Link counts	Sources documented

Graph 1: PBRF output versus Google Scholar citation count



Graph 2: PBRF output versus Google Scholar hits



References:

Bar-Ilan, J. (2005). Expectations versus reality – Search engine features needed for Web research at mid 2005. *Cybermetrics, 9*(1). http://www.cindoc.csic.es/cybermetrics/articles/v9i1p2.html

Bar-Ilan, J., & Lin, A. (2006). *Which h-index? - A comparison of WoS, Scopus and Google Scholar.* Paper presented at the 9th International Conference On Science And Technology Indicators, 7–9 September 2006 Leuven, Belgium.

Bauer, K., & Bakkalbasi, N. (2005). An Examination of Citation Counts in a New Scholarly Communication Environment. *D-Lib Magazine, 11*(9). http://www.dlib.org/dlib/september05/bauer/09bauer.html

Jacsó, P. (2005). Google Scholar: the pros and the cons. *Online Information Review, 29*(2), 208-214.

Kousha, K., & Thelwall, M. (2006). *Sources of Google Scholar citations outside the Science Citation Index: a comparison between four science disciplines.* Paper presented at the 9th International Conference On Science And Technology Indicators, 7–9 September 2006 Leuven, Belgium.

Morris Matthews, K., & Hall, C. (2006). *Impact of the Performance-Based Research Fund on teaching and the research-teaching balance: survey of a New Zealand University.* Paper presented at the Symposium on the evaluation of the PBRF, 16-17 February 2006, Wellington.

Noruzi, A. (2005). Google Scholar: The New Generation of Citation Indexes. *LIBRI, 55*(4), 170-180.

Smith, A. G., & Thelwall, M. (2002). Web Impact Factors for Australasian universities. *Scientometrics, 54*(3), 363–380.

Smith, A. G., & Thelwall, M. (2005). Web links as an indicator of research output: a comparison of NZ Tertiary Institution links with the Performance Based Research Funding assessment. In *ISSI 2005, 24-27 July, Stockholm.*

Tertiary Education Commission. (2004). *Overview and Key Findings: Performance-Based Research Fund: Evaluating Research Excellence: the 2003 assessment*. http://www.tec.govt.nz/downloads/a2z_publications/pbrf_report_overview.html

Thelwall, M. (2002). Methodologies for crawler based Web surveys. *Internet Research: Electronic Networking Applications and Policy, 12*(2), 124-138.

Thelwall, M. (2003a). A Free Database of University Web Links: Data Collection Issues. *Cybermetrics, 6/7*(1). http://www.cindoc.csic.es/cybermetrics/articles/v6i1p2.html

Thelwall, M. (2003b). What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation. *Information Research, 8*(3 paper no. 151). http://informationr.net/ir/8-3/paper151.html

Thelwall, M., & Harries, G. (2003). The connection between the research of a university and counts of links to its web pages: An investigation based upon a classification of the relationships of pages to the research of the host university. *Journal of the American Society for Information Science and Technology, 54*(7), 594-602.

Vaughan, L., & Shaw, D. (2006). *Comparison of citations from ISI, Google, and Google Scholar: seeking web indicators of impact.* Paper presented at the 9th International Conference On Science And Technology Indicators, 7–9 September 2006 Leuven, Belgium.

White, B. (2006). Examining the claims of Google Scholar as a serious information source. *New Zealand Library & Information Management Journal*, *50*(1), 11-24. http://www.lianza.org.nz/publications/journal/files/TNZLIMJOctober2006v50i01.pdf