Putting Google Scholar to the test: a preliminary study

Mary L. Robinson
St. Luke's Institute of Cancer Research, Dublin, Ireland, and
Judith Wusteman
School of Information and Library Studies, University College Dublin, Ireland

Received 7 April 2006
Revised 24 October 2006
Accepted 30 October 2006

Abstract

Purpose – To describe a small-scale quantitative evaluation of the scholarly information search engine, Google Scholar.

Design/methodology/approach – Google Scholar's ability to retrieve scholarly information was compared to that of three popular search engines: Ask.com, Google and Yahoo! Test queries were presented to all four search engines and the following measures were used to compare them: precision; Vaughan's Quality of Result Ranking; relative recall; and Vaughan's Ability to Retrieve Top Ranked Pages.

Findings – Significant differences were found in the ability to retrieve top ranked pages between Ask.com and Google and between Ask.com and Google Scholar for scientific queries. No other significant differences were found between the search engines. This may be due to the relatively small sample size of eight queries. Results suggest that, for scientific queries, Google Scholar has the highest precision, relative recall and Ability to Retrieve Top Ranked Pages. However, it achieved the lowest score for these three measures for non-scientific queries. The best overall score for all four measures was achieved by Google. Vaughan's Quality of Result Ranking found a significant correlation between Google and scientific queries.

Research limitations/implications – As with any search engine evaluation, the results pertain only to performance at the time of the study and must be considered in light of any subsequent changes in the search engine's configuration or functioning. Also, the relatively small sample size limits the scope of the study's findings.

Practical implications – These results suggest that, although Google Scholar may prove useful to those in scientific disciplines, further development is necessary if it is to be useful to the scholarly community in general.

Originality/value – This is a preliminary study in applying the accepted performance measures of precision and recall to Google Scholar. It provides information specialists and users with an objective evaluation of Google Scholar's abilities across both scientific and non-scientific disciplines and paves the way for a larger study.

Keywords Search engines, Information retrieval, Precision

Paper type Research paper

1. Introduction

Google Scholar (http://scholar.google.com) was launched in November 2004 with the aim of identifying scholarly information on the web and making it accessible. Its

The research described in this article was conducted by Mary L. Robinson in partial fulfilment of the requirements for the degree of Master in Library and Information Studies at University College Dublin.

Program: electronic library and information systems
Vol. 41 No. 1, 2007
pp. 71-80
© Emerald Group Publishing Limited
0033-0337
DOI 10.1108/00330330710724908

arrival prompted much discussion among both information professionals and the general scholarly community (e.g. Abram, 2005; Henderson, 2005; Jacsó, 2005a; Noël, 2005). A search engine for scholarly information, even a freely available one, is not a new idea. Two of the best known search engines of this type are Pennsylvania State University's Citeseer (http://citeseer.ist.psu.edu) and Elsevier's Scirus (www.scirus.com) (Sullivan, 2005). What is unique about Google Scholar is that it provides a means for accessing research in the humanities and arts, areas ignored by Scirus and Citeseer (Felter, 2005). However, Google Scholar's coverage is unknown as Google is reluctant to identify key aspects such as which publishers' articles are indexed, how scholarly information is identified, and how often its index is updated.

2. Evaluating search engines

The precision of a retrieval system is traditionally measured as the proportion of retrieved documents that are relevant, that is, number of relevant retrieved items/number of retrieved items. Recall, on the other hand, is the proportion of relevant documents that are retrieved i.e. no. of relevant items retrieved/no. of relevant items in the whole collection (Vaughan, 2004; Gordon and Pathak, 1999). Relative recall is the number of relevant documents retrieved by one retrieval system divided by the total number of unique, relevant documents retrieved by two or more retrieval systems; it is used where the total number of relevant documents is unknown. Because of the unique nature of the web environment, searching the web for information differs from searching in more traditional environments. Hence, alternative measures to precision and recall have been developed for evaluating web search services such as search engines. These alternatives include Vaughan's Quality of Result Ranking and Vaughan's Ability to Retrieve Top Ranked Pages (Vaughan, 2004). The Quality of Result Ranking measure, an alternative measure to precision, uses Spearman's Correlation to compare the relevance ranking assigned to each result by a search engine to the relevance ranking assigned by a user (Vaughan, 2004).

In the web environment, calculating a search engine's absolute recall is impossible because it requires knowledge of the total number of relevant documents on the web (Oppenheim *et al.*, 2000). Relative recall offers a practical, if imperfect, solution to this problem. It involves merging the relevant results returned by several search engines to arrive at an estimate of the number of relevant documents on the web. Relative recall is calculated as the number of relevant documents retrieved by a particular search engine divided by the number of unique relevant documents retrieved across the tested search engines.

Ability to Retrieve Top Ranked Pages is also based on the merged outputs of several search engines but differs from relative recall in that it relies on the ranking of results on a continuous scale. A search engine's Ability to Retrieve Top Ranked Pages is calculated from the number of documents returned by the search engine that are ranked by the user to be in the top 75 per cent or 50 per cent of merged results (Vaughan, 2004).

Views differ as to whether search engine effectiveness should be evaluated using real life queries or using queries generated by the experimenter for the purpose of evaluation (Borlund, 2000). Gordon and Pathak (1999) recommend the former; they also urge that the relevance of any returned results should be judged by the person requiring the information and not a third party such as the experimenter. Searches should capture, as fully as possible, the true needs of the searcher and should exploit the special features of each engine. Hence, identical search strings should not be presented to different search engines without due consideration.

3. Research questions and methodology

This study sought to investigate the following questions:

- Is Google Scholar better at finding scholarly information than other popular search engines?
- Does Google Scholar differ in its effectiveness in finding scientific and non-scientific scholarly information?

3.1 Search engines chosen

Searches using Google Scholar were compared to those using three other popular search engines: Google (www.google.com), Yahoo! (www.yahoo.com) and Ask.com (www.ask.com). These were chosen because they consistently appear in listings of the top three search service providers (Sullivan, 2005) and because they represent examples of three different types of search engine, namely free-text, index-based and natural-language respectively (Bradley, 2004). The inclusion of Google is important in order to identify how Google Scholar differs from its parent and because Google is particularly popular among academic users (Carroll, 2004). The index-based Yahoo! represents web pages by positioning them within a knowledge hierarchy developed and maintained by people, rather than by computer. The natural-language search engine, Ask.com, formerly Ask Jeeves, not only searches for specific terms but will also map them to other related terms (Bradley, 2004).

3.2 Search strategies

An Information Request Form, modified from Gordon and Pathak (1999), was used to elicit queries from seven postgraduate students. They were asked to describe their information needs, identify important phrases, words or synonyms, to phrase their search in the form of a Boolean query and to comment on their confidence that their query accurately captured the search question. As recommended by Gordon and Pathak (1999), each query was modified to take full advantage of the particular features of the individual search engines as described in their associated help pages. The queries were then presented to each of the four search engines. All searches were carried out over a six-day period from 24th to 29th July 2005. Searches for a particular query were run within a one hour period, thus minimising the potential for changes to the content or availability of websites.

3.3 Processing search results

As few users of search engines view more than 10 documents from a results list (Jansen and Pooch, 2001; Silverstein *et al.*, 1999), students were asked to rank only the top 10 result URLs returned by each search engine for their query. The URLs were presented to students via a Microsoft Excel spreadsheet, hence students did not know which search engines were being tested. The students were requested to identify each URL as either relevant or irrelevant and to rank each URL relative to all the search engine results for the same query. The entire study, from initial searches to relevance ranking by the students, was completed within a five-week period. The following four measures were used to compare the results: precision, Vaughan's Quality of Result Ranking, relative recall and Vaughan's Ability to Retrieve Top Ranked Pages.

All statistical analysis was carried out using the SPSS version 11 software (www.spss.com/). The Friedman F-test for a randomised block design (McClave *et al.*, 1997) was used to determine if there was a significant difference in any of the four measures between any of the search engines. Where significant differences were found at the .01 or .05 significance levels, a paired *t*-test (McClave *et al.*, 1997) was used to identify which specific search engines differed.

4. Results

4.1 Overview

Of the seven postgraduate students participating in this study, four were students of library and information studies and three were students of biology. One of the biology postgraduate students completed two Information Request Forms on two distinctly different topics with no overlap of results. Thus, a total of eight Information Request Forms were returned. The eight search queries were:

(1) (river AND geomorphology) AND (restoration OR rehabilitation OR sediment OR erosion);

(2) (freshwater AND macroinvertebrate) AND (recovery OR disturbance OR colonization OR niche OR extinction);

(3) ((insecticide AND environment AND impact AND study) AND (selective OR reduced-risk OR efficacy)) NOT (broad-spectrum);

(4) (algae AND spore AND substrate) AND (enteromorpha OR settlement OR microstructure OR biofouling OR colonization);

(5) (student AND international AND information AND library AND irish) AND (academic OR university OR third);

(6) ((traveller AND woman AND information AND irish) AND (need OR access OR service)) NOT (hotel OR flight OR holiday);

(7) (accessibility AND design AND html) AND (web OR internet OR wcag OR disability OR disabled); and

(8) ((metadata AND digital) AND (library OR libraries OR archive OR archives)) NOT (repositories OR repository), limited to previous 12 months.

As already mentioned, each query was modified to take advantage of the features of the individual search engines. The number of dead, out of date links and duplicated results are given in Table I. For simplicity, the name "Google Scholar" has been shortened to "Scholar" in all tables and figures.

For Query 7 (on Internet accessibility for people with disabilities), the top three results returned by Google Scholar were references to books. However, no links to further information were provided with these results and, as the only details given were the title and author, they could not be accurately evaluated for relevance. Hence, these three results were removed from the list sent to the student for relevance evaluation.

Only 11 results, or 1.13 per cent of all accessible results returned in these searches, were returned by more than one search engine as shown in Figure 1. This lack of overlap in search engine coverage is in agreement with that of previous studies (Lawrence and Giles, 1999; Jacsó, 2005b) and highlights the problems faced by users in attempting to search comprehensively for relevant information on the web.

In five of the eight searches, Google returned a link to an automatically generated Google Scholar search on the topic, along with the top three results. Google does not provide a link to Google Scholar from its generic search page; hence this may be Google's attempt to advertise Google Scholar's abilities to scholarly users.

4.2 Precision

4.2.1 Traditional precision measurements. As illustrated in Figure 2, the average precision across all queries ranged from 51.25 per cent for Google to 32.5 per cent for Ask.com. Google Scholar demonstrated the highest precision for scientific queries at 55 per cent, and Ask.com the lowest at 15 per cent. However, for non-scientific queries, Google and Yahoo! tied for the highest precision score at 52.5 per cent, while Google Scholar had the lowest precision at 30 per cent. However Friedman's F-test failed to find any significant differences in search engine precision at either the 0.05 or 0.01 level.

4.2.2 Vaughan's quality of result ranking. A significant correlation, at the 0.05 level, was found between Google's ranking and the student ranking of Query 2 and also, as shown in Table II, between Google and the student ranking of scientific queries in general. The value of Spearman's correlation coefficient (r) was 0.673 and 0.367

	Inaccessible results				Accessible results	
Query	Ask.com ($n = 10$)	Google ($n = 10$)	Scholar ($n = 10$)	Yahoo! ($n = 10$)	Total accessible ($n = 40$ − inaccessible)	Total duplicates
1	1	1	0	0	38	1
2	0	0	0	1	39	2
3	0	0	1	2	37	0
4	0	0	0	0	40	1
5	0	1	0	0	39	1
6	0	0	0	0	40	2
7	0	0	0	0	40	1
8	0	0	1	0	39	3
Total	1	2	2	3	312	11

Table I. Number of inaccessible, accessible and duplicated results returned by each search engine for each query

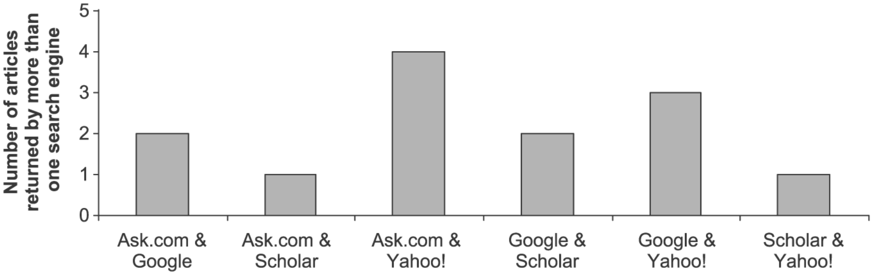

Figure 1. Overlap in search engine results

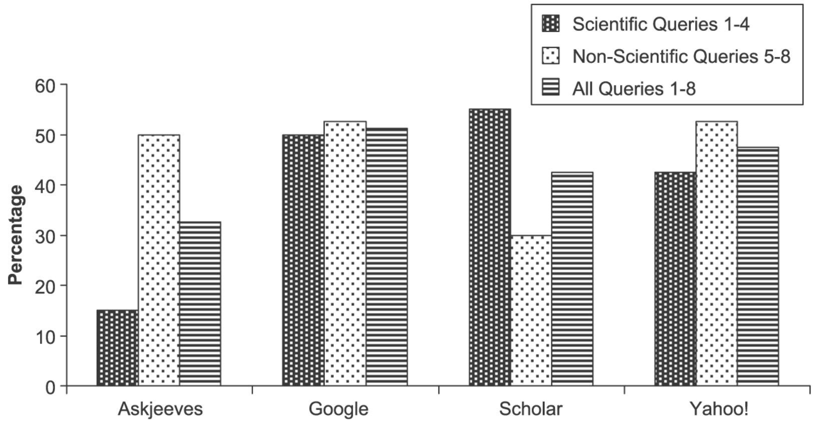

Figure 2.
Precision of search engines

Table II.
Quality of Result Ranking as measured by Spearman's correlation

	Scientific queries 1-4	Non-scientific queries 5-8	All queries 1-8
Ask.com	−0.236	0.1	−0.068
Google	0.367*	−0.182	0.092
Scholar	0.064	0.033	0.048
Yahoo!	0.014	−0.085	−0.036

Note: *Correlation is significant at the 0.05 level

respectively. A significant correlation was also found, at the 0.05 level, between Ask.com and the student ranking of Query 7 ($r = 0.685$). A significant correlation, at the 0.01 level, occurred between Yahoo! and Query 4 ($r = 0.806$), while a significant but negative correlation, at the 0.05 level, was found between Yahoo! and Query 3 ($r = -0.632$). Again, Friedman's F-test failed to find any significant differences in results returned by the search engines for scientific, non-scientific and all queries at either the 0.05 or 0.01 level.

The significant negative correlation between Yahoo! and Query 3 is surprising as it suggests that Yahoo! ranked the results in opposition to the student ranking. Vaughan (2004) found a similar significant but negative correlation between the ranking of the search engine Teoma (now merged with Ask.com) and the ranking assigned by users. Vaughan identified a problem with Teoma's phrase search function as the cause of its poor performance. No such problem could be found with the Yahoo! search functions in this study and the reason for this result remains unknown.

Although the differences in the precision and Quality of Result Ranking of the four search engines were not statistically significant, Google Scholar and Google achieved the highest scores in precision for scientific queries and across all queries respectively. However, Vaughan's Quality of Result Ranking found Google to be the best search engine for scientific queries. This apparent contradiction may highlight the importance played by a search engine's result ranking ability to bring relevant results to the attention of the user.

4.3 Recall

4.3.1 Relative recall. The total number of relevant documents returned across all queries was 130; 61 of these being returned for scientific queries and 69 for non-scientific queries. Figure 3 shows the relative recall values for these groups.

Friedman's F-test failed to find any significant differences in the relative recall of search engines across all queries. However, only documents that occur in the top 10 results were originally considered. It is possible that a document, appearing in the top 10 results of one search engine and rated as relevant, may appear outside the top 10 results of a second search engine. Hence, the second search engine would have a slightly higher recall than that calculated by simply considering the top 10 results it returned. To allow for this possibility, the top 100 results returned by each search engine were studied for any occurrences of these relevant documents. Thirteen such occurrences were found. These results were added to the original data and relative recall was recalculated. Again, Friedman's F-test failed to find any significant differences in the relative recall of search engines across all queries.

Although Google achieved the highest relative recall overall, the differences in the overall relative recall of the four search engines were not found to be statistically significant. Neither were significant differences found when relevant results appearing outside the top ten were included. However, Google Scholar achieved the highest relative recall for scientific queries and the lowest for non-scientific queries.

4.3.2 Vaughan's ability to retrieve top ranked pages. For each query, the students ranked the merged top 10 results returned by all four search engines on a continuous scale. The retrieval ability of each search engine was calculated as a percentage of the top 50 per cent of results returned, as illustrated in Figure 4.

Vaughan (2004) discovered that the choice between a cut-off point of 50 per cent or 75 per cent made little difference to test results. All relevant documents in this study were contained within the top 50 per cent of results for all queries except Query 7; for the latter, all relevant results were contained within the top 67 per cent. Hence, a cut-off point of 50 per cent was deemed reasonable.

At the 0.05 level, Friedman's F-test identified a significant difference in the ability of the search engines to retrieve relevant results for the scientific queries. At this same level, a paired *t*-test identified significant negative relationships between Ask.com and Google and between Ask.com and Google Scholar. As illustrated in Figure 4, Ask.com

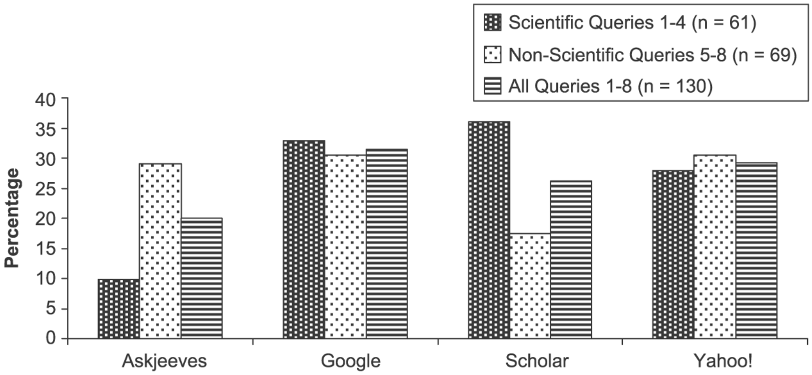

Figure 3.
Relative recall for scientific, non-scientific and all queries for all four search engines

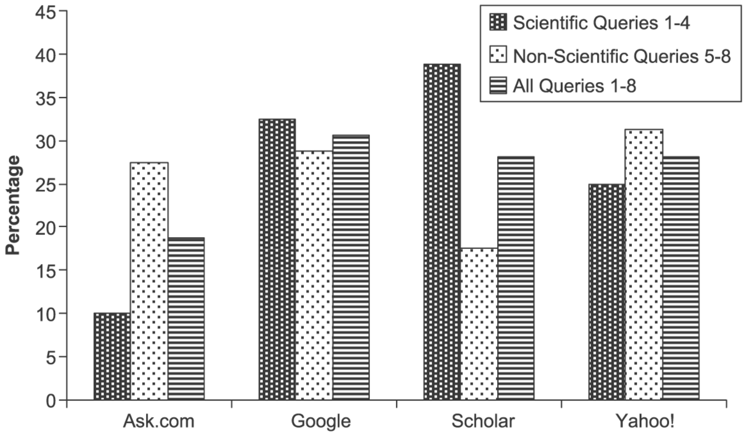

Figure 4.
Ability to retrieve top-ranked results

retrieved 10 per cent of relevant results; Google retrieved 32.5 per cent and Google Scholar, 38.75 per cent.

After relevant results appearing outside the top 10 were included for their respective search engines, the same calculations were repeated. Friedman's F-test identified a significant difference in search engine ability across total queries, scientific and non-scientific queries. Paired *t*-tests identified a significant difference, at the 0.01 level, in the ability to retrieve relevant results between Google and Ask.com. Again, this is supported by Figure 4 which shows that Ask.com gained a lower score than Google for all three groups.

Vaughan's Ability to Retrieve Top Ranked Pages identified a significant negative relationship, at the 0.05 level, between Ask.com and Google and between Ask.com and Google Scholar for scientific queries. This may be explained by Ask.com's poor performance in retrieving relevant scientific results relative to Google and Google Scholar.

5. Conclusions

The only significant differences in the ability to retrieve top ranked pages were found between Ask.com and Google and between Ask.com and Google Scholar for scientific queries. For scientific queries, Google Scholar had the highest precision, relative recall and Ability to Retrieve Top Ranked Pages. However, it achieved the lowest score for these measures for non-scientific queries. Vaughan's Quality of Result Ranking found a significant correlation between Google and scientific queries. The best overall score for all measures was achieved by Google.

The results of this small-scale study suggest that, before Google Scholar can fully achieve its goal of identifying and making accessible scholarly information on the web, its poor performance in relation to non-scientific queries must be addressed. Although it is not clear how Google Scholar identifies material as being of a scholarly quality, it is worth noting that many of the relevant results found for the non-scientific queries in this study were of a factual nature rather than of a scholarly nature. For example, Query 6 sought details on information services available to women in the Irish Traveller community. Whilst many of the resulting sites could aid scholarly research, they might not be described as scholarly *per se*. This may explain, in part, Google Scholar's poor performance in relation to the non-scientific queries.

Google Scholar fared better in retrieving scientific queries. But whether it is the best search engine for such queries remains to be determined. In a comparison with Scirus, Notess (2005) found Google Scholar to return fewer results. As the sources covered by each appear to be different, using both may be the best option.

This preliminary study suggests that Google Scholar offers potential benefits to both the novice and experienced scholar. Along with other free information retrieval services, it has an advantage over library-held resources in that students can continue to use it after graduation (Peek, 2005). Further, it offers a search experience familiar to anyone with even limited exposure to Google. Indeed, Felter (2005) suggests that Google Scholar may prove popular precisely because of its association with Google. In addition to supplementing services available from institutional libraries, Google Scholar may also prove a useful tool for those in teaching positions in checking for plagiarism, as pointed out by Zimmerman (Mayne, 2005).

Key questions about the Google Scholar search engine remain unanswered. For example, how long Google Scholar is expected to remain in Beta phase, what resources and publishers' records Google Scholar has access to and how often its index is updated (Tony Eklof Pers. Comm.). But no search engine indexes the entire web (Oppenheim *et al.*, 2000) and, given that search engines do not typically release figures on the freshness of their indexes, Google Scholar appears, at least, no worse than its peers in this respect. Indeed, these uncertainties may be of less concern to users than information specialists might fear. For example, Pyne *et al.* (1999) found that medical professionals rarely ask which journals are included in particular databases or how often those databases are updated. This highlights the vital role of information specialists in ensuring that users are aware of the abilities and limitations of scholarly retrieval services.

The findings of this initial study suggest that the Beta status of Google Scholar is still appropriate. Although already a serviceable tool for science scholars, the weaknesses discussed in this article must be addressed if Google Scholar is to prove useful to scholars of non-scientific disciplines.

References

Abram, S. (2005), "The Google opportunity", *Library Journal*, Vol. 130 No. 2, pp. 34-5.

Borlund, P. (2000), "Experimental components for the evaluation of interactive information retrieval systems", *Journal of Documentation*, Vol. 56 No. 1, pp. 71-90.

Bradley, P. (2004), *The Advanced Internet Searcher's Handbook*, 3rd ed., Facet, London.

Carroll, S. (2004), "Googled science", *LibraryConnect*, Vol. 2 No. 2, p. 5.

Felter, L. (2005), "Google Scholar, Scirus, and the scholarly search revolution", *Searcher*, Vol. 13 No. 2, pp. 43-9.

Gordon, M. and Pathak, M. (1999), "Finding information on the World Wide Web: the retrieval effectiveness of search engines", *Information Processing and Management*, Vol. 35 No. 2, pp. 141-80.

Henderson, J. (2005), "Google Scholar: a source for clinicians", *Canadian Medical Association Journal*, Vol. 172 No. 12, pp. 1549-50.

Jacsó, P. (2005a), "Google Scholar: the pros and the cons", *Online Information Review*, Vol. 29 No. 2, pp. 208-14.

Jacsó, P. (2005b), "Visualizing overlap and rank differences among web-wide search engines: some free tools and services", *Online Information Review*, Vol. 29 No. 5, pp. 554-60.

Jansen, B.J. and Pooch, U. (2001), "A review of web searching studies and a framework for future research", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 3, pp. 235-46.

Lawrence, S. and Giles, C.L. (1999), "Accessibility and distribution of information on the web", *Nature*, Vol. 400 No. 6740, pp. 107-9.

McClave, J.T., Dietrich, F.H. and Sincich, T. (1997), *Statistics*, 7th ed., Prentice-Hall, Upper Saddle River, NJ.

Mayne, P. (2005), "Good bye Britney Spears, hello academic journals", *Western News*, University of Western Ontario, London, 23 June, available at: http://communications.uwo.ca/western_news/story.html?listing_id = 18864

Noël, E. (2005), "Google Scholar", *Bulletin des Bibliotheques de France*, Vol. 50 No. 4, pp. 43-6.

Notess, G.R. (2005), "Scholarly web searching: Google Scholar and Scirus", *Online*, Vol. 29 No. 4, pp. 39-41.

Oppenheim, C., Morris, A., McKnight, C. and Lowley, S. (2000), "The evaluation of WWW search engines", *Journal of Documentation*, Vol. 56 No. 2, pp. 190-211.

Peek, R. (2005), "A Googly New Year", *Information Today*, Vol. 22 No. 1, pp. 17-18.

Pyne, T., Newman, K., Leigh, S., Cowling, A. and Rounce, K. (1999), "Meeting the information needs of clinicians for the practice of evidence-based healthcare", *Health Libraries Review*, Vol. 16 No. 1, pp. 3-14.

Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. (1999), "Analysis of a very large Web search engine query log", *ACM SIGIR Forum*, Vol. 33 No. 1, pp. 6-12.

Sullivan, D. (2005), *5th Annual Search Engine Watch Awards*, available at: http://searchenginewatch.com/awards/article.php/3494141

Vaughan, L. (2004), "New measurements for search engine evaluation proposed and tested", *Information Processing and Management*, Vol. 40 No. 4, pp. 677-91.

About the authors

Mary L. Robinson is Assistant Librarian at St. Luke's Institute of Cancer Research, Dublin, Ireland.

Judith Wusteman is a Lecturer in the School of Information and Library Studies, University College Dublin, Ireland. She is the corresponding author and can be contacted at: judith.wusteman@ucd.ie