Googling South African academic publications - search query generation methods

M Weideman
Website Attributes Research Centre
Cape Peninsula University of Technology
PO Box 652
Cape Town
0027 21 9135515
weidemanm@cput.ac.za

ABSTRACT

Commercial websites need to rank well on search engines to ensure a high degree of exposure. This has become true even for academic webpages. Many universities store research outputs in institutional repositories. However, free-form Internet searching is still preferred by many students. This implies that academic publications should be findable via standard search engines. The purpose of this research was to determine the best type of query to lead to a high-ranking website containing the required abstract when searching for academic publications.

A questionnaire was used to gather data on published research outputs. A variety of search queries were constructed for each output, and these queries were run to find research output abstracts online.

An analysis of the results provided a variety of patterns when searching for known published academic content. The patterns were different for journal articles, conference papers, books and theses.

Some of the abstracts were ranked highly on Google search result pages, with others not ranked among the first ten results. Overall the best results were provided by combining the author surname with the first sentence of the abstract text.

It was concluded that the generation of search queries be alternated between the two most successful query types, rather than focussing on one type only.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *query formulation, search process*

General Terms

Measurement, Documentation, Human Factors.

Keywords

query generation; academic publication; search engines

1. INTRODUCTION

Prior research has shown that most website owners put a high premium on ranking well on SERPs (search engine result pages) for given keywords and key phrases. This is especially true if the website is based on selling of a commercial product or service. The higher a website ranks on a SERP, the more traffic it will draw through user clicks [16]. As a result, there is a strong element of competition in the commercial world for ranking well for a given keyword or key phrase. Common terms such as "London hotel", "books for sale" and "house insurance" produce respectively 20.6 million, 65.2 million and 89.1 million search results when run on Bing. These high figures are an indication of the oversupply of websites with relevant answers, again confirming the competitive nature of the process of achieving high rankings.

Search engine marketing involves spending money on either one or both of an SEO (search engine optimisation) or a PPC (pay per click) approach. Both are aimed at improving the ranking of a website on a SERP, although it is implemented in different ways. Although much more advertising dollars are spent on PPC, most users still prefer clicking on SEO-based results [11].

Finding the full-text of academic papers on the Internet without payment is an ongoing battle, as described in the literature on open access. For the purpose of this research, it was assumed that the first sentence of the abstract and other basics (author surname, etc) are known. It is further assumed that the user will search only using free-form searching. Free-form Internet searching is defined as the use of standard search engines (eg. Bing, Google and Yahoo!) for finding resources, as opposed to academic databases.

The purpose of this research was determining the best method of generating Internet search queries for finding academic abstracts through free-form Internet searching.

2. PRIOR RESEARCH

2.1 Internet searching

Since the birth of the Internet and its subsequent phenomenal growth, users have become used to resorting to search services for finding relevant information. Generally users have shown an increase in searching skills by the increase over years in the length of search queries used [12]. However, success in information retrieval depends also on whether or not the required information is visible to search engine crawlers, and indeed, indexed at all. A 2009 study attempted to determine what the impact and visibility was of a series of publications across various disciplines [13]. It was found that the visibility, access to full-text articles and thus the impact was quite low. Ford *et al* did a study to explore the impact of differences between human individuals on the Web searching behaviour [6], and determined that there are sizeable differences in the way users approach searching. Other authors [9] studied information visibility, in which an attempt was made to better understand searcher behaviour.

A recent study on the performance of natural language search engines vs standard ones, proved that there was not much difference in precision between these and Google. It appears as if Internet searching using these two kinds of products produce similar results, thereby removing the apparent advantage of natural language searching [7].

In summary, Internet search engines do not provide a fool-proof method of finding relevant information. Human behavior differs widely, interfaces sometimes impede progress and lack of visibility of indexed information could prevent available information from being found.

2.2 Open access

A major trend in academic publishing during the last decade has been the move towards open access - academic publications which are freely available as opposed to being accessible through expensive subscription services. Generally the responsibility of paying for the publication is moving towards the author(s) of the publications [10]. One major contributor to this debate was the Finch Report on open access [5].

A study was done to determine what the impact of open access journal articles in four widely varying fields of research have on research in general [1]. Measurements were done in terms of ISI Web of Science citations, and it was found that (across all disciplines), articles that were freely available had a greater research impact than those which were not.

The librarian's attitude to open access was studied in a project on UK librarians, and it was found that a suspicion of open access materials was prevalent, coupled with researchers being unaware of institutional policies with regards to open access [2].

So, although open access presents exciting opportunities for information consumers, there are still many stumbling blocks before it can be successfully implemented. Not least of all is the sheer cost - this author has been offered publication of a peer reviewed journal article under open access for a fee of €2150 [4].

2.3 Search query generations

The creation of effective search queries has been identified as one of the most important stumbling blocks on the way to successful Internet information retrieval. A system was developed to automatically generate a query when searching for patents, in an attempt to overcome this difficulty [17]. At least two USA patents have been registered towards solving this problem:

- Patent No 6,564,213 B1, May 13, 2003 (Search Query Autocompletion), and
- Patent No 6,772,150 B1, Aug 3, 2004 (Search Query Refinement using Related Search Phrases).

Another author suggested an approach to query generation based on a sliding scale finding the equilibrium between a too short (too many answers) and too long (too few answers) search query [16]. See Figure 1. The same author claims that a "basic formula" for building a search query could lead to searching success:

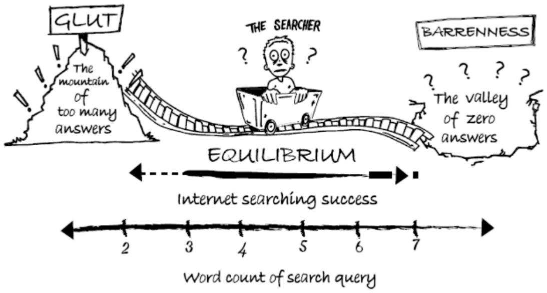

GLUT
THE SEARCHER
BARRENNESS
The mountain of too many answers
The valley of zero answers
EQUILIBRIUM
Internet searching success
2
3
4
5
6
7
Word count of search query

Figure 1. In search of the perfect search query. [16]

"Express the information need as a single, keyword-rich English sentence.
- Remove all stop words from this sentence.
- Type the remaining string of keywords into a search engine search box."
[16].

2.4 Ranking

It has been proven that a website with value to its owner must rank highly on a SERP for a given set of keywords/key-phrases. A number of studies have been done on the ranking of websites, and how to improve ranking. Evans [3] used 50 competition websites as a starting point, and found that a variety of factors influence ranking, including a webpages' PageRank, inlinks, directories and social bookmarks. Weideman produced a prioritized ranking model in 2009, claiming that inlinks and body keywords are the most important ranking factors [16].

Sullivan followed this up with a practitioner-oriented model, arranging contributing factors on a periodic table-type layout [14]. Killoran did another study in 2013 to determine how SEO techniques can be used to increase website visibility [8]. It was noted that website content should be closely matched to the intended audience, and the importance of using relevant keywords in body content was emphasized.

For the sake of clarity, the following terminology will be used in this research when referring to the ranking of a website on a SERP. A *high ranking* will mean a numerically high figure (i.e. 10, 20, etc), which is a "bad" ranking. A *low ranking* will refer to numerically low figure (i.e. 1 or 2), which is a "good" ranking.

3. METHODOLOGY

3.1 Questionnaire

A questionnaire was designed to extract information about four types of completed and published research outputs from South African academics: journal articles, conference papers, books/book chapters and theses. For each type of output, enough information was requested to enable this author to construct a variety of search queries. These were needed to establish whether or not the abstract of the given output was accessible via free-form searching on the Internet. Authors were numbered randomly, and only these numbers are used in the figures following to preserve anonymity of participants.

The following information was requested for each output type:
- the title,
- the year of publication,
- the surnames of all the authors,
- the full abstract of the publication, and
- three keywords/key phrases describing the content.

Prior research has proven that various combinations of weight-carrying keywords often provide effective search queries [15]. A decision was therefore taken to generate three different queries for every search, using the information requested, as described below.

- Search Query a (called Qa hereafter): The respondent's surname was combined with the first sentence of the abstract.
- Search Query b (called Qb hereafter): The year of the publication was combined with the surnames of all the authors and all the keywords.
- Search Query c (called Qc hereafter): The first sentence of the abstract was combined with all the keywords.

Examples of each search query type follow, where dummy surnames were used to protect the identity of participants. In each case the character **X** is used to delimit the sections as noted above

Qa:
Anderson **X** Gas samples were taken from a wide range of target areas on dumps arising from coal mining activities

Qb:
2008 **X** Anderson **X** Brown **X** Cooper **X** Semantic Web **X** Semantic Web Architecture **X** Usefulness of Semantic Web Architecture

Qc:
The purpose of this study is to develop a business framework for the effective start-up and operation of African immigrant businesses in the Cape Town Metropolitan Area of South Africa **X** business framework **X** start-up **X** African immigrant

4. RESULTS AND DISCUSSION

4.1 Results

A total of 26 participants responded and completed the questionnaire, providing information on a total of 51 journal articles, conference papers, books and theses. Not all authors submitted information for all four types of outputs, as expected. For example, only 15 of the 26 participants provided information on published journal articles.

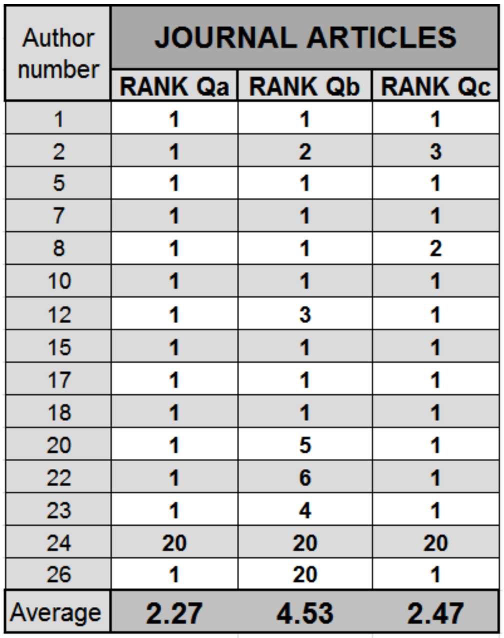
The values used in the body cells of Table 1 to 4 are an indication of the actual ranking achieved by that result which leads the searcher to a copy of the required abstract. The figure 1 for instance, indicates the best ranking possible, where the abstract was found at the website listed first on the SERP. When an abstract was not found inside the first 10 results (i.e. on the first SERP), an indication had to be given that this ranking was worse than those in the top 10. This is the case since the website containing the abstract was either ranked worse than position 10, or not indexed at all. If a zero or an empty space were to be used, it would cause the average figure to present a more positive picture than the real situation - it would decrease the average figure instead of increasing it. Therefore the value 20 was entered into all empty cells, since it would increase the numerical value (i.e. worsen the ranking) of the average figure.

For every output type, an average was calculated. Again, a lower average figure indicates a better ranking for that specific query generation method.

4.2 Journal articles

A total of 15 out of the 26 participants did list a journal article as part of their answer set - see Table 1.

Table 1. Rankings for journal article searches.



Author number	JOURNAL ARTICLES		
	RANK Qa	RANK Qb	RANK Qc
1	1	1	1
2	1	2	3
5	1	1	1
7	1	1	1
8	1	1	2
10	1	1	1
12	1	3	1
15	1	1	1
17	1	1	1
18	1	1	1
20	1	5	1
22	1	6	1
23	1	4	1
24	20	20	20
26	1	20	1
Average	2.27	4.53	2.47

For one author (no 24) a copy of the abstract could not be found amongst the top ten results on the SERP. Qa produced the best possible result, with 14 first positions out of 15 possible ranks. The average figures confirm this - Qa has the lowest (best) rank of 2.27, followed by Qc with 2.47 and last Qb with 4.53. The average of these three rankings is 3.09.
These figures indicate that Qa provides the most and Qb the least effective query generation methods for finding academic abstracts on the Internet.

4.3 Conference papers

Fourteen of the 26 participants supplied information for published conference papers, as presented in Table 2.

This time a total of six of these 14 conference paper abstracts could not be found on the first SERP. Qa again produced the best possible result, with the lowest average ranking of 7.5, followed by Qc with 10.5 and lastly Qb with 17.08.

Table 2. Rankings for conference paper searches.

Author number	CONFERENCE PAPERS		
	RANK Qa	RANK Qb	RANK Qc
1	1	20	1
2	20	20	20
5	1	20	20
7	1	20	1
8	1	20	1
10	20	20	20
12	20	20	20
15	1	3	1
17	20	20	20
18	1	20	1
20	2	2	1
22	2	20	20
23	20	20	20
24	20	20	20
Average	7.50	17.08	10.50

The average of these three rankings is 11.69. Compared to the previous set of results conference papers have a much lower visibility and worse ranking than journal articles.

4.4 Books/book chapters

The lowest number of participants (11 out of 26) provided details for books/book chapters, and theses (Section 4.5). Table 3 provides the summary for books/book chapters.

Table 3. Rankings for books/book chapter searches.

Author number	BOOK/BOOK CHAPTERS		
	RANK Qa	RANK Qb	RANK Qc
1	1	1	1
2	1	1	1
5	20	20	20
8	1	1	1
10	1	1	1
15	3	20	3
17	1	1	1
20	20	20	20
22	20	20	20
24	20	20	20
26	20	20	20
Average	4.00	6.43	4.00

Five of the 11 book/book chapter abstracts could not be found on the first SERP. This is the highest percentage so far, indicating that book/book chapter abstracts are less prevalent on the Internet and less accessible by free-form searchers than journal articles and conference papers. Qa and Qc produced identical results of an average ranking of 4.00, with Qb again performing the worst with an average of 6.43.

The average of these three rankings is 4.81.Compared to the previous set of results books/book chapters have a visibility just slightly worse than that of journal articles, but slightly better than those of conference papers.

4.5 Theses

From the 11 respondents, four were not ranked on the first page at all - see Table 4. The average of the figures is 9.24.

Table 4. Rankings for theses searches.

Author number	THESES		
	RANK Qa	RANK Qb	RANK Qc
1	1	1	1
2	1	20	1
8	1	2	1
10	1	1	1
11	1	3	1
12	20	20	20
15	20	20	20
17	20	20	20
19	1	4	1
20	20	20	20
24	20	1	1
Average	9.64	10.18	7.91

The pattern established by the previous three types of outputs, in terms of search query effectiveness, has been, without exception:
Qa most effective, then Qc, then Qb being the least effective.

However, considering the theses results, a reversal in the first two positions is evident - Qc being the most effective with 7.91, followed by Qa with 9.64 and lastly Qb with 10.18.

Qb has consistently performed the worst with a third place in all four cases.

5. CONCLUSION

Finally, the performance of the search query generation methods are compared - see Table 5 and Figure 2.

Table 5. Average rankings: method vs output type.

	JA	CP	BC	TH
Qa	2.27	7.50	4.00	9.64
Qb	4.53	17.08	6.43	10.18
Qc	2.47	10.50	4.00	7.91

The four types of outputs are listed on the horizontal axis: JA for journal articles, CP for conference papers, BC for books/book chapters and TH for theses.

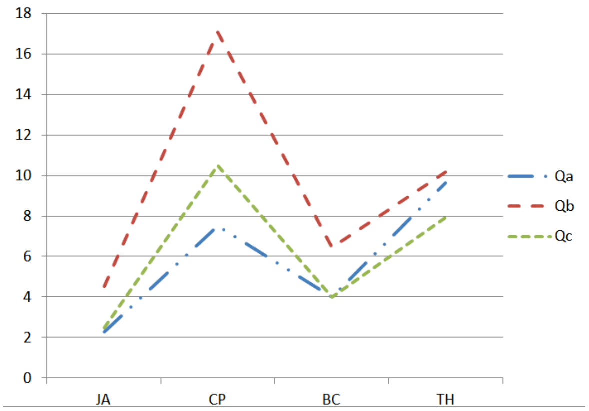


Figure 2. Average rankings for three query generation types.

It is clear that the graph for Qb consistently has the highest point value for each of the four types of outputs, indicating that it has produced the worst ranking.

It can be concluded that combining the year of publication with author surnames with keywords is the least effective query generation method to be used when using free-form searching for academic publications.

In an attempt to quantify the positions of the other two methods, their point values (Table 5) and graphs (Figure 2) need to be compared. Qa is more effective than Qc in two of the four cases, equal in the third and worse in the fourth. Calculating the average rankings proves that Qa is the best method overall with an average rank of 5.85, followed by Qc with an average of 6.22. Qb is the least reliable method with an average of 9.56. However, the difference between Qa and Qc is small.

In final conclusion, the common factor between the queries generated for Qa and Qc is the inclusion of the first sentence of the abstract in the search query. In a well-written academic abstract, the first sentence should be keyword-rich, descriptive and indicative of the topic of the academic document. This research has proven the value of weight-carrying keywords as part of a query generation strategy.

It can thus be concluded that the method identified in this research bears close resemblance to the "basic formula" for searching identified in Section 2.3. Searchers are advised to generate search queries by combining descriptive, weight-carrying keywords, which could include surnames of authors and sentences from abstracts, for most effective academic information retrieval.

References

[1] Antelman, K. 2004. Do Open Access Articles Have a Greater Research Impact? College & Research Libraries News, 65, 5, 372-382.

[2] Creaser, C. 2010. Open access to research outputs—institutional policies and researchers' views: results from two complementary surveys. New Rev Acad Libr. 16, 1 (Apr), 4–25.

[3] Evans, M.P. 2007. Analysing Google rankings through search engine optimisation data. Internet Research, 17, 1, 21-37. DOI 10.1108/10662240710730470

[4] Feinstein, J. (jessica.feinstein@tandf.co.uk). 2013. *RE: Open access is available for your article*. E-mail to M. Weideman (weidemanm@cput.ac.za). 30 May 2013.

[5] Finch, J. 2012. Accessibility, sustainability, excellence: how to expand access to research publications. http://www.researchinfonet.org/wp-content/uploads/2012/06/Finch-Group-report-FINAL-VERSION.pdf

[6] Ford, N., Eaglestone, B., Madden, A. and Whittle, M. 2008. Web searching by the "general public": an individual differences perspective. Journal of Documentation, 65, 4, 632 - 667.

[7] Hariri, N. 2013. Do natural language search engines really understand what users want?: A comparative study on three natural language search engines and Google. Online Information Review, 37, 2, 287 - 303.

[8] Killoran, J.B. 2013. How to use search engine optimization techniques to increase website visibility. IEEE Transactions on Professional Communication. 56, 1, March.

[9] Mansourian, Y., Ford, N., Webber, S. and Madden, A. 2008. An integrative model of "information visibility" and "information seeking" on the web. Program: electronic library and information systems. 42, 4, 402-417.

[10] McCabe, M.J. and Snyder, C.M. 2005. Academic Journal Quality. AEA Papers and Proceedings, May.

[11] Neethling, R. 2008. Search engine optimisation or paid placement systems - user preference. Masters Thesis, Cape Peninsula University of Technology, Cape Town.

[12] Phan, N., Bailey, P. and Wilkinson, R. 2007. Understanding the relationship of information need specificity to search query length. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '07), New York, USA. July 10 - 13. 709-710. doi>10.1145/1277741.1277870

[13] Russell, J.M., Ainsworth, S. and Diaz-Aguilar, J. 2012. Web visibility or wasted opportunity? Case studies from Mexican research institutes. Aslib Proceedings, 64, 1, 67 - 82.

[14] Sullivan, D. 2011. Introducing: The Periodic Table of SEO Ranking Factors. http://searchengineland.com/introducing-the-periodic-table-of-seo-ranking-factors-77181

[15] Weideman, M. 2010. Empirical study on crawler visibility of PDF documents in digital libraries. In Proceedings of The Third IEEE International Conference on Computer Science and Information Technology (ICCSIT 2010), Chengdu, China. July 10 - 13. 373 - 379.

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

[17] Xue, X. and Croft, W.B. 2009. Automatic query generation for patent search. In Proceedings of the 18th ACM Conference on Information and Knowledge management (CIKM '09), Hong Kong, China, November 2 -6, 2037-2040