The Impact of Search Engines on the Quality of Research by Tertiary Students

Completed as part of the Masters of Information Management program at RMIT, Melbourne, Australia

Submitted 27 August 2010

Ben Chadwick

Table of Contents

Introduction

The impact of the internet and related technologies on scholarly research is an ongoing issue in contemporary academic and information-specialist arenas (Grafton 2009). One particularly prominent issue is the use of commercial internet search-engines for scholarly research, and how this bears upon the utilization of traditional and ostensibly more rigorous information-retrieval systems (IRs) online public access catalogues (OPACs) and internet databases. The discourse surrounding this topic frames the use of search engines (SEs) – and the most popular engine, Google, in particular – as either pathological mutations of the research process, or the next miraculous stage in its evolution. Williams (2007) suggests that we "marvel at [Google's] speed, flexibility and simplicity", and urges that "it is adaptation... which ensures the survival of a species". By contrast, Haigh (2006, p. 33) proclaims that the "Google Society [is] convinced it is on the verge of a bright, shiny, networked utopia... even as it reduces its culture to machine-generated lists of what everyone else is looking at, so stupid that it does not realize how stupid it is".

The current paper seeks to address the issue of the impact of SEs on scholarly research, with reference to the comparative role of library-mediated IRs. The scope of the current paper will emphasize two topics that feature in contemporary research in the area: the research activities of university students, and the SE Google.

It will begin by describing Google and the SEs, and the issues of concern associated with Google, many of which are applicable to all SEs. Then, the

performance of SEs as scholarly information resources will be discussed. Next, the extent of SE usage for scholarly research will be explored. Finally, the nature of user's interactions will be examined in terms of their potentially negative impact on scholarly research.

Google and the SEs

Xie (2004) discusses two types of IRs that are suited to differing informational needs. Traditional IRs, including bibliographic databases and OPACs, are generally available through institutional libraries. They have more complex search features and are generally suited to topics of a non-personal nature. The content of these IRs is highly selective and managed, and derives an authoritativeness from a network that extends from the credentials of authors, review boards and journal editors, publishers of journals, editors of databases, and the screening and selection activities of librarians (Regalado 2007; Doldi & Bratengeyer 2005). Libraries subscribe to bibliographic databases for public use, and although they are still available in off-line modes, such as CD-ROM, modern databases are usually available online[1].

Popular IRs include SEs.[2]. Google, introduced in 1998, is the most popular SE, with a 25% extra market share over its nearest competitor, Yahoo! (Sullivan

---

[1] In this paper the term *traditional IRs* will be used to refer to all traditional, library-mediated IRs as described here. For the sake of simplicity, when the term *databases* is used it will be in reference online bibliographic databases.

[2] In the current paper, when Google is not being discussed explicitly, the term *search engines* will be used to refer to popular IRs as described here.

2006). Its success it partly attributable to the innovative PageRank algorithm for determining the 'relevance' of search results (Haigh 2006) by calculating the number of links to a page, akin to citations. PageRank also assigns weights to the referring pages on the basis of factors such as history and institutional significance (Chen, Kraemer, & Sharma 2009; Haigh 2006).

Perhaps the concern expressed most readily by authors in the field is over the potentially poor quality of resources returned by search-engines (Regalado 2007; Haigh 2006; Williams 2007). Haigh (2006) raises a number of basic operational concerns about Google. The retrieved by PageRank are *popular*, but this doesn't guarantee they are *relevant* to a research question. Google are not forthcoming about the workings of their algorithm, making it impossible to fully understand the mechanisms underlying a search. In addition, the PageRank system is self-reinforcing – the sites that are returned are more likely to be further cited, and therefore even more likely to be returned in future searches. Thus, Google searches only a surface of the internet, creating the phenomenon of the 'dark-web' where certain resources, including large academic and government databases, are never searched.

Haigh (2006) also notes the influences that come to bear on a Google search that may undermine research integrity. The results of a Google search are readily influenced both deliberately, such as by search-engine optimization, or inadvertently, such as by blogs or discussion sites that contain multiple redundant links to a site. Furthermore, commercial 'sponsored links' are returned with each search, and although they are displayed separately to the general results, two-thirds of users do not discriminate between the two sources.

Performance of SEs

Concerns such as these cause reservations about Google's use as an IR for scholarly material, particularly given the availability of traditional IRs that, by virtue of their design, would ostensibly direct users to higher quality information. By contrast, Google's authority derives from the effectiveness of PageRank's algorithms at ranking results according to popularity. As Williams (2007) notes, information that is easy to find isn't necessarily the best.

To understand how SEs compare to databases as IRs for scholarly information we will review three empirical studies that directly compare the performance of each in locating resources for specific scholarly queries.

Firstly, in terms of the total number of results returned, the two studies reporting this data found that SEs returned between 1.25 (Doldi and Bratengeyer 2005) and 2.5 (Brophy and Bawden 2005) more results than traditional IRs.

Findings relating to the relevance of results were more inconsistent, but it appears that the source judged to return the most relevant results depends on who judges 'relevance'. When judged by novices, SEs tended to perform better (Xie 2004). In the case of Brophy and Bawden (2005), who judged relevance themselves but adopted strict criteria and used a larger 'sample' of search topics, the two sets of IRs performed more comparably. Doldi & Bratengeyer (2005) also judged relevance themselves but did not describe their methodology. In their case the results swung towards what might be considered the direction of an experimenter's predilections: traditional IRs outperformed search-engines.

Brophy and Bawden (2005) examined the number of records returned by Google and traditional IRs that were not returned by the other system ('uniques'), and found that Google was superior, finding roughly 1.66 as many uniques as traditional IRs.

Only one study compared IRs on the quality of results, and found that traditional IRs were superior, returning 1.1 times more high-quality results than Google (Brophy and Bawden 2005).

Doldi & Bratengeyer (2005) also found that whilst databases returned no duplicates, one-quarter to one-third of SE results were duplicates.

Brophy and Bawden (2005) examined the accessibility of original documents, and found that Google returned twice as many relevant items that were immediately retrievable[3].

Finally, on a more qualitative level, Doldi and Bratengeyer (2005) noted that results from databases were consistently well-structured, presented uniformly, and contained abstracts. By contrast, web-based results did not conform to any format and may not have had abstracts, meaning that web-results may take longer to process and their relevance may not be readily ascertainable.

The interpretation of the above findings is ambiguous. It could be argued that because SEs return far more results than traditional IRs (after all, they are free to trawl the entire visible web rather than a selection of records), they produce

---

[3] However, the authors noted that if the Google result was not immediately accessible then it was not accessible at all, whereas library system results that were not immediately accessible could, in most cases, be located "with difficulty" via linked resources, bringing the total percentage of locatable items to 93%.

more results that are relevant but of lesser quality. Such resources, including white-papers, research briefings, and blogs and opinions that are *unique* to search-engine results, are also generally more readily available to download.

Utilization of SEs

Grafton (2009) states that the contemporary student's "primary source of information on life, the universe, and everything is the Web" (p. 96). It might be suggested that the potential shortcomings of SEs could be overlooked if only they were not relied upon for scholarly research. However, the largest survey to date of SE and library-resource usage reveals the opposite.

In 2005 the Online Computer Library Center (OCLC) surveyed 3348 individuals from Australia, Singapore, India, Canada, the UK, and the US. Both the general population and college students reported that they had used SEs (68% and 75% respectively) more than online libraries (26% and 47% respectively) and were more familiar with SEs (36% and 45% respectively) than online libraries (8% and 20%, respectively). Even more striking was participant's indications of where they would begin an information search, with SE usage (84% and 89% respectively) far exceeding online libraries (1% and 2% respectively).

Research training and experience appears to alter these preferences: studies of upper-level humanities students (Head 2007) and postgraduates (Riahinia & Zandian 2008) demonstrate a preference for traditional IRs over SEs.

The Relationship Between SEs and Users

Inherent in Haigh's (2006) article is a recognition that the shortcomings of Google itself belie the more fundamental issue of how search-engines are used. Regaldo (2007) describes a generational effect resulting from the coalescence of parallel socio-technological trends over recent decades. Because the do-it-yourself nature of internet publication allowed individuals to produce and distribute material whilst bypassing traditional editorial processes and gate-keeping mechanisms, a vast array of materials became available to students that could be accessed without reliance upon traditional information-seeking authorities. Also, whilst the authoritativeness of these materials varied widely, for many the authority vested in traditionally published materials transferred to web-based materials. Finally, rules surrounding the authority of information seekers changed: students were no longer dependent on librarians or reading lists, but possessed their own authority as primary users of new technologies.
Far from being 'blank slates', Grafton (2009) observes that the ‘Google-generation’ of students come to universities with highly developed information-seeking styles resulting from exposure to the internet, SEs, and sites such as Wikipedia. Why would a student adopt new information-seeking skills when their current style has been adequate for their needs to date, and the IR tools they use remain most convenient and accessible?
Haigh (2006, p31) describes a survey of 2316 US academics, in which 42% reported that they believed internet usage impacted negatively on the quality of student's work. Grafton (2009, p. 96) states "students normally seek information not by making a research plan but by entering words in a search engine - usually

a non-specialist one". Students' then perceive rankings on a results-page as an indication of authority (Regalado 2007).

As shown earlier, these results probably include a wide array of probably-relevant, but not necessarily high-quality resources.

Even what is returned may not be browsed adequately – Jansen and Spink (2006) found that 73% to 76% of internet searches did not go beyond the first page. Furthermore, links from a results-page may be broken and electronic resources may not be accessible, or, students may simply choose to use the 'snippets' available on the results page as their information source. These 'snippets' are disembodied pieces of information, free from a meaning-generating context: on a results-page it is not often clear who the author is or with what authority they write; but users tend to trust the credibility of this content nonetheless (Haigh 2006; Regalado 2007; Williams 2007). Grafton (2009) suggests that because texts are not consumed as coherent wholes, but skimmed and picked-at, it is not possible to form a critical understanding of the structure of the text. Students scan in a focused and goal-oriented manner on isolated *pieces* of information rather than considering a text or problem in its wider context. Grafton states that students "think they are making effective, critical use of materials of every kind, which are in fact torn from the context that is vital to critical judgment" (p. 96).

When original web-pages or electronic documents are accessed, large amounts of data can be downloaded, contributing to an overburden of potentially irrelevant information, and a tendency to put reading off – perhaps indefinitely (Haigh 2006).

Perhaps more importantly, Haigh (2006) suggests that the research methods of 'Google-generation' students limit the degree to which even source materials are processed. Grafton (2009) claims these students do not study web-sites or electronic documents intensively but skim them and move on to the next, spending an average of four minutes on an e-journal and eight minutes on an e-book. If a link doesn't appear to immediately satisfy an information need users can easily move on to a new link, meaning sources are poorly scrutinized. Also, with the ease of cutting and pasting, materials can be selected and utilized with very little processing of content (Haigh 2006). Related to this, of course, educators claim the prevalence of cut-and-paste plagiarism has increased with use of internet sources (Grafton 2009, Regalado 2007).

A lack of context can be an issue even for original documents and web-pages found via SEs. Williams (2007) says, "it is much easier to find a document using Google than to identify its origin, the parties responsible for it, or whether the information it contains is correct" (p6). This state of affairs contrasts strongly with traditional monographs and journal articles, where the credentials and experience of the publisher, editors, and author are clearly defined (Williams 2007).

Conclusion

It may be a little melodramatic for Haigh (2006, p33) to proclaim that "Google-society" is "reducing its culture" to the point that it "does not realize how stupid it is". As a scholarly IR, Google may not be ideal, and an over-reliance upon it may foster less-than-ideal research methods. However, it might be argued that the

information-seeking heuristics of the Google-generation are superior to those of previous generations, which may have involved little more that following the popular press and absorbing the opinions of friends and family. For some informational needs, Google works very well, and for more authoritative material we have seen that it works adequately much of the time. Surely, for the 'novice' population, this is better than practicing entirely naive or highly problematic information-seeking strategies.

Perhaps the greater challenge facing educators today is not the existence of Google *per se*, but, as Williams (2007) put it, "persuading [students] not to stick to the shallows of what is readily available to them".

If information-seekers are not exposed to traditional IRs they will never have the opportunity to learn to use them effectively, to experience their benefits, and to observe some of the shortcomings of web-based systems and the research-heuristics associated with them. It is only via this process that they will cease being the 'novices' and become the experts.

References

Brophy, J & Bawden D 2005, Is Google enough? Comparison of an internet search engine with academic library resources, *Aslib Proceedings: New Information Perspectives*, vol. 57, no. 6, pp. 498-512.

Chen R, Kraemer, KL & Sharma, P 2009, Google: The world's first information utility? *Business and Information Systems Engineering*, vol. 1, pp. 53 -61.

Doldi, LM, & Bratengeyer, E 2005, The web as a free source for scientific information: A comparison with fee-based databases, *Online Information Review*, vol. 294, pp. 400-411.

Grafton, A 2009, Apocalypse in the stacks? The research library in the age of Google. *Daedalus*, vol. 138, no. 1, pp. 87- 98.

Haigh, G 2006, Information idol: How Google is making us stupid, *The Monthly*, February, no. 9, pp. 25-33.

Head, AJ 2007, Beyond Google: How do students conduct academic research? *First Monday*, August, vol. 12, no. 8 – 6, viewed 2 August 2010, http://firstmonday,org/htbin/cgiwrap/bin/ojs/index,php/fm/article/viewArticle/1998/1873.

Jansen, BJ, Spink, A 2006, How are we searching the www: A comparison of nine search engine transaction logs, *Information Processing and Management*, vol. 421, pp. 248-263.

Online Computer Library Center 2005, College Students' Perceptions of Libraries and Information Resources, OCLC, Dublin, Ohio.

Regalado, M 2007, Research authority in the age of Google, *Library Philosophy and Practice*, LPP Special Issue on Libraries and Google, viewed 22 August 2010, http://www.webpages.uidaho.edu/~mbolin/regalado.pdf.

Riahinia, R & Zandian, F 2008, Evaluation of information providers and popular search engines on the base of postgraduate students' perspectives, *The Electronic Library*, vol 26, no. 4, pp. 594-604.

Sullivan, D 2006, Nielsen NetRatings search engine ratings, viewed 24 August 2010, http://searchenginewatch.com/2156451.

Williams, G 2007, Unclear on the context: Refocusing on information literacy's evaluative component in the age of Google, *Library Philosophy and Practice*, LPP Special Issue on Libraries and Google, viewed 23 August 2010, http://www.webpages.uidaho.edu/~mbolin/williams.htm.

Xie, H 2004, Online IR system evaluation: Online databases versus Web search engines. *Online Information Review*, vol. 28, no. 3 , pp. 211-219.