Aspects of Google: bigger is better - or less is more?

Purpose of this paper	To investigate recent enhancements to the internet search engine Google.
Design/methodology/approach	An opinion piece based on practitioner experience and recent commentary on search engine innovations.
Findings	That recent innovations in Google's functionality have yet to deliver what they promise, but that it is still to early to say what can genuinely be achieved in these areas.
Research limitations/implications	This is an expression of opinion about a service that will be radically improved and developed in the immediate future.
Practical implications	Gives some useful insights and tips on how to use existing digital library tools to achieve information retrieval results along the lines of those aspired to by Google.
What is original/value of the paper?	An attempt to give clear, practice-based examples of how to apply recent digital information retrieval developments to contemporary library work.

Keywords: search engines; digital libraries.

Introduction

Google is widely acknowledged as the world's favourite internet search engine. However, it has ambitions to better itself – which is typical of Google's dynamism and inventiveness and is to be applauded. So, two important changes have happened to Google: firstly, it has been expanding its scope, in order to embrace more of the so-called 'invisible web' (a form of scaling up, on the basis that bigger is better). Secondly, it has also tried to create a more focussed version of itself, Google Scholar[1], in order to exclude the irrelevant and trivial aspects of the web, and to provide a gateway into academic and weighty information (a form of narrowing down, implying that less is more). Google's plan to link up with Stanford, Michigan and Oxford University Libraries, in order to digitise significant parts of their paper-based scholarly collections obviously complements this aim to create a dedicated scholarly search engine.

These are both worthy plans aimed at enhancing the quality of an already high quality product, but they are not without their pitfalls. Let us reflect on what these might be.

Google and the invisible web

As for Google and the invisible web, I've already written an editorial in Library Review vol. 53 no. 7 about some of the problems this expansion of searching can cause.

The 'invisible web' often comprises, among other things, databases full of metadata. But one of the traditional strengths of the web (and it's interesting that the web is now old enough to have 'traditional' strengths!) is the preponderance of full-text material which it retrieves. I once asked a Business student why he didn't use the library's online periodical indexes to find material for an essay. The reply was, 'Once I've found something in a periodical database, I've just got a reference. So then I've got to find it on the library shelves. And maybe it's not there anyway. If I search Google, I'm probably going to get a whole set of web pages with lots of text about the company names I've put into it. I want text not a lot references which may or may not be on the shelf. '

You can't argue with customer preference (although that was a while back, and there are now better full-text links in online periodical databases). However, it is at least arguable that this shows that Google users have always liked its full-text retrieval in preference to the metadata of library databases. It's ironic then that in trying to improve its scope by entering the invisible web, Google has starting pillaging repositories of metadata. This is not necessarily playing to its strengths.

My previous example in vol. 53 of Library Review was of searching on Google for full-text details of houses on an estate agent's web site, and being deluged by metadata from database entries about the estate agent and its website, rather than the site itself. Here is another example of the problem, which may serve as a useful illustration.

Our reference desk staff were asked by a Politics student for any information about an independent Socialist candidate who stood in the last Scottish parliamentary election called 'something like' Jimmy McNaughton. In particular, the user wanted to know which constituency he had stood in. Now my very own institution hosts an excellent web site about that particular election. The site is a product of the Aspect project, and forms part of the Glasgow Digital Library[2].

So the next appropriate step might have been to go to the GDL web site and browse through it to get to the relevant entry. But in fact (rightly or wrongly!) our next step was to throw the key words 'Glasgow Aspect McNaughton' into the Google search box, expecting it would get pretty close to the relevant page on the Aspect web site.

You can try this quick and dirty search for yourself (assuming the content and search algorithms of Google haven't changed too much since we tried this search). It in fact retrieves a record[3] from a database - the previously invisible web - about the hitherto unreported problem of people sitting on decrepit sanitary ware, which, due to ageing, collapses under their weight,

This is a peer-reviewed author accepted manuscript of the following research article: Joint, N. (2005), "Aspects of Google: bigger is better – or less is more?", Library Review, Vol. 54 No. 3, pp. 145-148. https://doi.org/10.1108/00242530510588890

thus creating injuries to the unfortunate individuals. It is an unexpected false drop (no pun intended), and certainly not an 'aspect' of 'Glasgow' that should come top of anyone's relevance ranked search list. But it is a result of a search engine poking into a database such as Pubmed[4], which has its own very well-developed metadata scheme designed to be searched in a tailored way by its own search engine. It is unsurprising that an unhelpful result was retrieved. Try the same keywords in Pubmed itself and you get no such false drop.

So, in as much as this possibly unrepresentative search tells us anything, it shows that expanding searches from the visible web to the invisible web does not always add value. In terms of search engine coverage, often 'less is more', rather than 'bigger is better'. In particular, if a search engine is going to start expanding its scope and trawling collections which have their own detailed metadata schemes, there is little point in doing so if:

a) the enquirer wants 'data' and not 'metadata'
b) the metadata is not used properly by the search engine (that is, the metadata is not used by the internet search engine to the same effect and in the same way as the local database search engine)

In the end we browsed through the relevant web site and found the information on the correct web page[5]. Unsurprisingly, the original query was slightly incorrect (the candidate's name was McNaught not McNaughton), but the information was easy to tidy up and locate. It was a librarian's brain, not Google, that came up with the answer in the end, which is quite reassuring for those of us who still run reference desks staffed by real people.

Google Scholar

In contrast to the idea of Google scaling up to trawl the invisible web, the creation of Google Scholar seems to represent the reverse, the process of focussing down on a particular collection of data. Not all of the Google world, but a narrower part of it. Given the difficulties of this internet search engine's hijacking myriad other databases' metadata, the idea of taking less information and searching it better seems compelling.

However, it is difficult to know whether this is exactly what Google Scholar offers us. Google Scholar is intended to trawl the metadata of specifically scholarly information services – so the invisible web is hardly excluded, just the non-scholarly bits of it. So one may assume that, if Google (G) represents all the web and Google Scholar (GS) the scholarly part of it, then, in Boolean terms, GS is a small circle contained entirely with the larger circle of G. Logically therefore, it should be easier to find information in response to the reasonably scholarly Scottish Political reference query above. However, my own efforts to reproduce either the correct information or the bizarre false drop quoted above proved unsuccessful. Does this mean that my Boolean metaphor is incorrect, and that the small circle GS only partly overlaps with the larger G, rather than

being subsumed by it (though surely there is *some* significant overlap between the two circles)? Or maybe I'm using the search engine wrongly (is there an information literacy librarian available to help please?).

Well, you can go on playing the game of trial searches ad infinitum, and some commentators have recently done this to good effect and in greater depth[6]. However, the message that emerges is the same:

a) 'Content is the most obscure part of Google Scholar' (ibid.) – you don't really know what is or is not contained within the service.
b) Without a firm grasp of the size and subject content of a collection (be it digital or hardcopy), relative to other benchmark collections, it's difficult to know what to do with a searchable collection of data.

Much of the effort of information literacy classes is in fact directed at guiding naïve users through the discrete datasets available to them via their home institution's electronic services - not only teaching them how to use different interfaces of each service, hitherto the main thrust of much IL training, but rather imparting a sense of the individual character and relative size of each digital collection, as in point b) above. And now portal technology offers the chance to aggregate these services and make them searchable in a large single sweep via a single interface.

At present, this seems a more fruitful way of getting the best of both worlds – 'bigger is better' (a single overarching portal search) or 'less is more' (isolate a single service and use it for a focussed search within a smaller dataset with a defined subject character). Google isn't in the running – as yet. Whether it will be soon is an open question. But given the quality of its service to date, one would be foolish to think its initial attempts to break new ground in this regard are anything other than the first steps in an exciting exploration of new possibilities. What the fruits of this exploration will be remains to be seen, but my hunch is that there is much better to come.

Nicholas Joint,
Editor,
'Library Review'

Notes and references

1. Google Scholar. < http://scholar.google.com/ > (last accessed 17th December 2004).

2. Glasgow Digital Library: Aspect Project
< http://gdl.cdlr.strath.ac.uk/aspect/ > (last accessed 17th December 2004).

3. Wyatt JP, McNaughton GW, Tullett WM. (1993) The collapse of toilets in Glasgow. *Scottish Medical Journal,* Vol. 38 no. 6 p. 185.

4. Pubmed < http://www.ncbi.nlm.nih.gov/PubMed/ > (last accessed 17th December 2004).

5. Aspect Project list of candidates.
< http://gdl.cdlr.strath.ac.uk/aspect/candidates/index.html >
(last accessed 17th December 2004).

6. Péter's Digital Reference Shelf. Review of Google Scholar Beta.
< http://www.galegroup.com/free_resources/reference/peter/dec.htm > (last accessed 17th December 2004).

This is a peer-reviewed author accepted manuscript of the following research article: Joint, N. (2005), "Aspects of Google: bigger is better – or less is more?", Library Review, Vol. 54 No. 3, pp. 145-148. https://doi.org/10.1108/00242530510588890