Meeting Report: Agenda for Information Retrieval

ISKO UK Meeting 26 June 2008, University College, London

Things have changed in more ways than one in the world of information retrieval. Such a glib assessment does little justice to the remarkably well attended meeting of the ISKO at University College, with getting on for 100 participants. Part of the success was due, no doubt, as chair Stella Dextre Clark pointed out, to the impressive range of speakers, starting with a very sprightly Brian Vickery, aged 90, about his career in information retrieval. It was sobering to think that this man has spent more time in retirement keeping up with information retrieval than most people spend on it in their working lives. But it is in our working lives that information retrieval has changed from a specialist discipline to a topic of central importance to most of our lives. This was illustrated clearly in Stephen Robertson's presentation, when he pointed out that he no longer needs to explain what he does – nowadays, everyone has some idea of what searching is all about.

When commentators describe the present-day information retrieval landscape, it is difficult not to appear trite, because so much of what is said is common knowledge, even if awareness of what search engines actually do is superficial. Google is a general search engine, of course, but it was Brian Vickery who pointed out that most retrieval systems over time have been unlike Google, and built with a specific user community in mind.

Vickery described the process by which communities of practice, in a specific domain such as chemistry, created knowledge that was private, and only part of that knowledge was subsequently transmitted to the public domain for retrieval. The implication, not picked up by the later speakers, was that the creation of a universe of information is in itself partial and inaccurate.

Of course, there is always the approach to information management that you should keep it simple – Vickery recalled the head of the British Library lending division, who stored all the books as they arrived alphabetically by title, without using any classification system. Well, it was certainly simple.

Stephen Robertson, currently based at the Microsoft Research Laboratory in Cambridge,, gave a fascinating presentation about types of information retrieval, using for much of his presentation the examples over twenty years or more of research into IR at Cranfield University, notably the TREC competition. Stephen provided an excellent analogy of how the world has changed because of Google: it is not unreasonable, he said, to describe Tim Berners-Lee as the inventor of the World Wide Web in the 20th century, but the Web in the 21st century is largely the creation of Google.

He then gave a very rapid review of the information retrieval process, which, while being non-technical, was enough to make you realise just what a complicated activity searching is. His simple example of searching for one word, then two words, then a phrase such as "black hole", where the meaning of the phrase is different to the meaning of either of the two component words, showed just why search engine companies need teams of search strategists.

As for the search heuristics, a simplistic assessment of all the years of research is that statistical methods of search always win when compared with other methods, such as directory-based or natural-language processing. This doesn't mean the other methods are superseded; they can frequently be used as an adjunct to improve searching. Many of these ideas have been around for years, for example a user adding comments to the result list (either the hits they see as relevant, or those they see as not relevant) to make subsequent searches more precise. Stephen's conclusion was that there are opportunities within enterprise searching to make some advances on the standard, one-size-fits-all Google model.

Finally, Ian Rowlands gave a fascinating description of how what has been dubbed the "Google Generation Report" came about, and some of his personal views on it. This report, which appeared in 2007, was commissioned by JISC and the British Library. Its goal was to determine if there was a specific Google generation (those born after 1993) who have grown up with no knowledge of a world without search engines and the Web.

Perhaps not surprisingly, the report concluded there was no major difference between generations – instead, there were transgenerational sectors, groups of people with similar characteristics whatever their age. The Google generation was found to comprise , perhaps unsurprisingly, a high-tech group, happy with leading-edge innovation, plus a majority of "average Joes", not technical, but using current technology; and finally a substantial number of "digital dissidents" (including Rowlands' own daughter) who consciously rejected many of the latest technical innovations and awareness of information technology. In other words, some "silver surfers" are more members of the Google generation than many 17-year-olds.

He pointed out that where work needed to be done is in discovering how people used digital resources. Some of the researchers at the CIBER centre at University College have discovered remarkable activities of online resource searchers, such as users of Elsevier Science Direct, who go into the service, find a single page, following a single search, then leave. Either these people are information marksmen, or they aren't finding at all what they want and giving up.

A further worrying piece of research was the Superbook project, 2007. A set of 3,000 e-books were added to UCL Library, and a study was made of how those books were retrieved. First, a diagram of the navigation process to reach those books showed the route to the information to comprise many difference routs, most of them remarkably convoluted, and often resulting in dead-ends. Secondly, he showed that the most popular route to the e-books was via the library catalogue, rather than Google.

A remarkable study carried out by Florida State University into student levels of information literacy revealed a correlation between information literacy skills and examination performance. Not so surprising, you would think, but the researchers then asked the students how confident they were of their information literacy skills. The top quartile of IL results rightly thought they had good IL skills, but it was worrying to find that the bottom quartile thought too that they had good information literacy skills. A further revelation from a detailed study of web logs was how little time many users spent actually reading electronic documents – there was a lot of skimming, but little detailed reading.

After presenting these dramatic research results, Ian's conclusions were for me perhaps a little disappointing. He talked about the need to promote information literacy from an early age, something easier said than done, and talked about the need to make information retrieval more appealing – perhaps a compulsory component of academic courses. His presentation concluded with a more personal interpretation of these results. He recollected as a school student doing his homework in Plymouth Central Reference Library, and having a powerful sense of the physical information structure with a collection of books – learning and knowing how and where information was to be found. This skill, he said, has to a large extent been lost. Hence the Google Generation report was in a sense asking the wrong question. Instead, we should be asking "How do you maintain information literacy in an environment that lacks the clues provided by a traditional library environment?"

My own conclusion is rather different. People have always read in very different ways depending on the material, their goals, and the nature of the text. The only difference is that today we have better tools to be able to track the process of retrieval and reading almost down to the specific words retrieved by users when reading a text. I don't think reading has changed, simply the tools for recording it. Roland Barthes pointed out (in *The Pleasure of the Text*) how we don't read every word of novels we are excited by, and it is common knowledge by researchers into the psychology of reading that reading is almost never word by word – the act of reading varies enormously depending on the reader's prior knowledge, goals, and context.

The question-and-answer session after the presentation revealed considerable concern from members of the audience, usually parents, about the presumed lack of knowledge of information retrieval shown by children (usually their own children). So perhaps the only Google generation is a generation of parents who worry about their children's education – and that is nothing new.

Michael Upshall