Discovery and Discoverability: New Ideas, Techniques and Products

Michael Upshall, Consult MU
Michael@consultmu.co.uk

This was the latest in a series of annual meetings organised by University College London's Centre for Publishing and held at Senate House's Chancellor's Hall, (a magnificent room with its original Art Deco fittings) on January 18th, 2017. The half-day conference comprised six presentations, a combination of research (in the first half) with applications (in the second). As often happens with events of this kind, there was little coherence about the presentations; the researchers made some interesting points, and the products described were interesting, but there was little by way of a grand theme to pull the event together. This is not a criticism of John Akeroyd, of University College London, who has managed this event for some years. As usual, he assembled an interesting line-up of speakers; but it reflects quite accurately a world of searching where there is little consensus about what constitutes best practice. Nonetheless, one or two themes emerged from the talks: these were serendipity and taxonomy.

First, the theory. Mounia Lalmas from Yahoo! stated boldly that "algorithms are not enough". That was a blow to those of us in the audience who had just discovered that in the machine-learning world, algorithms are the answer to all known problems. Professor Lalmas described how Yahoo was building discovery algorithms to engage users, as well as researching how users behaved. Her definition of user engagement was being "captivated by the technology", a phrase that some people might question, and included engagement by feelings and interactions, as well as by serendipity.

Professor Lalmas had the advantage of a team and expert knowledge of research procedures to investigate user engagement; but unfortunately, the conclusions she presented were less than startling. Comparing users of Wikipedia with users of Yahoo! Answers, she concluded that Yahoo! Answers delivered more unexpected results; and that "interestingness" is not equal to relevance. A study of usage in Chile suggested that different cultures (she contrasted metropolitan and rural areas) had different notions of relevance.

More directly relevant to academic search was Emine Yilmaz's talk, bravely entitled "New Developments in Search". Dr Yilmaz proposed the undeniable truth that all searches for a topic are actually motivated by an underlying task. We may search for a mortgage, but we really want to buy a house. She then presented a few alarming-looking mathematical equations on the screen, but I suspect I was not alone in responding more to the non-mathematical findings she presented in support of the need to think about tasks rather than topics. One screen showed the difference in time taken by users when they carry out common web-based activities. For example, the task of "planning travel" typically involves twelve minutes per travel activity, but web-based searching took only five

minutes of this – that is, less than 50% of the total time involved in the planning activity. In other words, measuring the way people search using the Web does not tell the whole story.

The challenge of her research is how we find out what users' tasks really are. All we have are search logs, so while we can all agree that next-generation search tools should be task-based, how in practice can we implement this? Our task is more difficult because, as she pointed out, people frequently carry out multiple tasks at once, or drift to and from different tasks while searching. She described organising the user tasks in a Bayesian rose tree structure, but I confess I didn't quite understand how she managed to capture user tasks before assembling them as a rose tree (even though I loved the idea of all my search tasks captured in this form). In conclusion, her first recommendation for the future was learning how to extract tasks, so I think she perhaps agrees that this approach requires some method of identifying tasks - unless perhaps we interview every user every time they search.

The next talk was my own, so I have to declare an interest. I described UNSILO, a tool that uses machine-learning to extract concepts automatically from texts. UNSILO is one of a new generation of machine-learning tools that work by statistical analysis – putting it rather simplistically, the engine looks at millions of words of text in a subject domain and identifies which phrases are significant for each document (these are the "concepts"). Behind the scenes, it carries out a lot of natural language processing and semantic analysis, so it is more semantic matching than string matching. Once the concepts have been identified, the system can identify related documents, or identify trending topics, or suggest a relevant journal, and many other functions.

Although UNSILO can use existing taxonomies, it does not require a pre-existing taxonomy to work. The talk questioned if we really need taxonomies in the first place; taxonomies add another stage to the content discovery process and can significantly add to the cost and time required to index content. At least, it seems, since several machine-learning tools can identify and categorise content without using taxonomies (Yewno, presented later that afternoon, is similar in that it requires no taxonomy) perhaps we should be questioning the entire taxonomy-based approach to discovery we currently use as the answer to all our search problems.

The next presentation was a refreshingly open and honest one by Timothy Hill, an engineer on Europeana, the website that aggregates cultural heritage content (or in most cases metadata about content). Europeana now comprises over fifty million objects, and Dr Hill described some experiments to try to enhance the Europeana data using semantic tools. I couldn't help feeling that while the approach was impressive in describing quite openly the routes the team had tried – including ones that turned out not to be so valuable – my overall impression was that the rather mixed results were not due to limitations of search so much as the problems of the source material and the site goal. Firstly, Europeana comprises metadata in over thirty different languages, with only partial translations available. Secondly, and perhaps even more fundamentally, Europeana is a collection of heritage information, but the use case for such a collection is not entirely clear - or at least, not reducible to one or two use cases. People will come to Europeana

for a vast range of purposes, but it is perhaps not as easy to identify simple use cases such as the standard Google or Yahoo!-type information use case, where the user wants to find a local restaurant, for example. Analysis of searches found that 70 - 95% of searches were for entities, such as "Rembrandt", but this of course does not reveal the underlying goal of the user.

Dr Hill's conclusions were not so surprising – enhancement works best when you have good metadata to begin with, using consistent spelling. More interesting was the observation that applying a standard taxonomy across several domains proved problematic: terms in one domain, such as place names, proved to have a different (and unintended) meaning when appearing in another domain. This suggests a further limitation of taxonomies; that they are very domain-specific and often cause problems outside their intended domain. More revealingly, the Europeana work revealed that linking to resources such as DBpedia (the machine-readable version of Wikipedia) was often better suited to cultural heritage collections than using formal taxonomies.

Incidentally, Dr Hill revealed on the theme of serendipity, that some Europeana users state explicitly that they come to the site to find something new – something that they didn't know before starting their search. Well, that's clearly a serendipitous aim, but quite how you could measure the extent to which you have satisfied user requirements in this case is mystifying.

The afternoon ended with two demonstrations of tools aiding discovery. "PowerTagging", from Digirati, combines a full-scale content management system (UMBRACO), and a full-scale taxonomy editor package, PoolParty, so that users can tag new content and edit their taxonomy at the same time. This is the kind of approach that suggests that if discovery is difficult, then we should all become taxonomists. In this case, the software did provide the user with a way to interact with the machine-created tags: the user, in this case the in-house system operator, can select or deselect concepts that match or don't seem appropriate. A drawback is that new terms for the taxonomy have to be inserted at the right place in the hierarchy – not for the faint-hearted. It reminded me of Heather Hedden's book *The Accidental Taxonomist*, which starts from the recognition that very few people get involved in indexing and classification by choice.

Finally, there was a presentation of Yewno, another machine-learning based discovery tool. Yewno's business case is pitched at institutions. It provides a visual, graph-based discovery service that searches across content from many publishers (now totalling some one hundred million items). Yewno does not hold the content, simply the concepts. Users can browse via the visual interface to identify topics that are matched to specific content items, which can be journal articles, book chapters, and so on. The content itself is held by the institution where the user is searching so this looks to be largely an institutional researcher tool, since without access to the content, Yewno would be a rather partial experience. The presentation concluded with a demonstration, which of course included several serendipitous results.

The event ended with a panel session with questions from the floor, and a couple of questions seemed to catch the presenters off-guard. One question was: "if using humans

to measure results is so difficult, why do we try to carry out human-based measurement?" The other question asked simply; if most researchers in practice use Google Scholar for their initial academic searches, why not just continue to use Google? This question, right at the end of the day, raised an issue that had not been discussed earlier in the meeting: the challenge, unrelated to search, of access to content, which is restricted by rights management. No matter how clever your search tools are, if you aren't searching all the possible content in academic search, you can't be certain you have found the correct answer. Google Scholar represents probably the largest collection of searchable academic content available, where both open-access and subscription content is included, because all commercial publishers make their content available to it. As a result, Google Scholar will always be the starting point of choice for many researchers – not because it is the best (there is an amusing blog post pointing out its limitations), but because it is the biggest. In contrast, Science Direct, or Web of Science, will only ever include a proportion (perhaps 50%, but still only a proportion) of all available academic content. In other words, however clever software tools might become at improving discovery, the reality is that more searches will continue to be made using Google Scholar than any other tool. And we have no control over the quality of search in Google Scholar. That's a rather sobering conclusion for an afternoon spent looking at discovery tools.

Michael Upshall has been involved in content enrichment for several years. He is currently head of business development for the Danish machine-learning company UNSILO.