On retrieval system theory

Stephen Robertson
Microsoft Research Cambridge, UK

Abstract

This paper re-reviews Vickery's book *On retrieval system theory*, first published 50 years ago, and discusses the changing nature of theoretical work on information retrieval and the possibility of developing a general theory of IR.

Introduction

Brian Vickery's book *On retrieval system theory* (Vickery, 1961) first appeared a half-century ago. (My copy, bought in 1967 when I took the City University masters course in Information Science, is the second edition of 1965.) Mooers' original phrase 'information storage and retrieval', shortened to 'information retrieval', had been coined in 1950, to represent a topic and problem area which was just then being recognised as significant – Vickery's further shortening to 'retrieval' on its own is interesting. At any rate, it represents a very early view of the field of information retrieval. Inevitably, in many respects it looks somewhat dated – and of course, Brian's view developed and changed hugely over the rest of his life. Nevertheless, it is of interest to examine what the field looked like, through his eyes at the time.

One obvious difference concerns mechanisation, automation, computerisation. Vickery comments in his introduction that 'both traditional and machine methods of indexing' are becoming more complex, and that the two 'are only variant ways of dealing with the same problem'. But one has to remind oneself that what he meant by 'machine' was not really computers – for further discussion see below.[1]

Vickery credits John O'Connor with having suggested the first word of his title (it's an example I have taken very much to heart – a significant number of my own papers use the same first title word). He also comments in his introduction 'There is as yet no unified theory of retrieval systems'. This is also something I have said many times myself, and continue to say. He goes on to say '... and a good deal of retrieval practice is still an empirical art, unsullied by theory.'

I suppose that what both of us seem to imply by such statements is that there will one day be a unified theory, or at the least that this is a desirable end to aim at. However, I will be suggesting in this paper that perhaps the non-existence of a unified theory is not only necessary but even

[1] All quotations from the book reproduced by permission of Reed Elsevier (UK) Limited, trading as LexisNexis.

desirable – that we can (and maybe should) revel in the mongrel nature of the field of information retrieval.

Science *versus* technology

In the modern world, we tend to associate science and technology together. Arguably, however, they are not only very different, but in a fundamental sense, opposed to each other. My ancient Shorter Oxford Dictionary defines technology as 'a discourse or treatise on an art or arts; the scientific study of the practical or industrial arts, practical arts collectively'. This is a revealing definition, containing four occurrences of the word *art/arts* and only one of *science/scientific*. The fundamental difference is that science is about understanding the way the world is, and why it is thus; technology is about changing the world, making it other than it is now. The greatest achievement that a scientist can claim is the discovery of a law, which (in science as much as in society) is about limiting the possibilities – about identifying some of the worlds that may be conceivable but cannot actually exist. The greatest achievement of a technologist is to succeed in doing something that might have been thought impossible.

In this sense, information retrieval is a technology rather than a science. There may of course be fundamental laws, basic theoretical limits to what is possible in the way of information retrieval, and we would certainly like to know what they are. But whenever any of us theorists (I regard myself primarily as a theorist) puts forward a candidate for such a law, it actually constitutes a challenge to the technologists amongst us: find a way around this limitation. If we can bring in an idea, a method, a technique, an approach from left field, as it were, to solve an apparently insoluble problem, then of course we will do so.

To some extent, I think the history of information retrieval as an empirical art over the last half century has demonstrated this principle to an extraordinary degree, and nowhere more so than in the present generation of general-purpose web search engines. Arguably other domains of information retrieval (such as intranet search) have lagged behind. This is because the technologists (working under extraordinary economic motivating forces) have discovered ways of doing things which work astonishingly well on the web, but may simply not be applicable elsewhere. It doesn't really matter if you are one of those people who thinks that Google is magic, or if you are frequently infuriated by its failure to guess what you really meant – either way, you have to marvel at how different it is from anything you or Brian Vickery might have imagined in 1961 or even 1986.

Communication sciences

In his first short chapter on 'The Scope of this Study', Vickery discusses the three nouns of his title. The third part, on Theory, starts by equating this word with theoretical generalisation, in other words with describing general principles of design and operation, rather than with particular systems. He goes on to appeal to the aid of many 'tool subjects', which he characterises as forming the broad area of Communication Sciences. He quotes an overview of this field in the form of a classification[2] drawn up by a pair of MIT scientists. This starts with Mathematical Foundations and

[2] As is entirely appropriate, given that Vickery's previous book was *Classification and Indexing in Science*.

Communication Processes, the latter subdivided into Non-living systems, Living systems, and Dynamics of large systems. The Living systems section, in which one might hope to find any human concerns, has Linguistics, Neurophysiology, Experimental psychology and Group behaviour. To me at least, this feels somewhat mechanically oriented. Vickery does go on to warn about applying ideas from these other subjects to retrieval without first establishing a good match between the form of communication that takes place in retrieval and the models used in these other subjects.

Some of the subjects referred to have their own body of theory, but others look themselves more like empirical arts. Certainly it is hard to discern any basis here for a unified theory. Also the analysis seems to miss out some relatively theoretical areas that might be regarded as 'home grown', on which Vickery himself had already done substantial work – the obvious example being the theory of classification and knowledge organisation. But the main point is that we in the information retrieval field, seeking both theory and praxis, may need to expand our own theoretical ideas by cherry-picking from other fields, as seems appropriate and useful, rather than seeking a single coherent theory which will both explain everything we need to explain and guide us in creating whatever we need to create.

Vickery's notions of generalisation and abstraction are further developed in a later chapter on automation, see below.

Vickery's chapters

The second chapter, 'The Analysis of Retrieval Systems', outlines the components of a retrieval system, and describes in delightful style a kind of system which has been 'in operation for some time', namely the MARLIS, ('multi-aspect relevance linkage'). This is a beautiful Brian Vickery joke, a description of a traditional library as if it were a modern machine system (including its computing unit, the HOMO).[3]

The central five chapters of the book are as follows: The Description of Documents, Descriptor Languages, Structural Models, File Organization and Coding, and Search Procedures. It's worth a very brief comment about the File Organization and Coding chapter, before putting it aside: an early consideration in this chapter is 'term entry versus item entry'. This distinction was transmuted (in the early stages of computer-based searching in the 1970s) into that between serial search and inverted files, an argument which was largely won for inverted files, in all except the tiniest of document collections (if your mail client's search function is still only capable of serial search on your mailbox, you probably never use it!). Otherwise this chapter is all but incomprehensible today, being about how to encode short complex subject descriptions of documents into forms which can be searched by a human searcher, in a catalogue or index.

The other four present an interesting bias, by today's standards at least. A large part of the ensuing half century's research might be seen as simplifying document description, ignoring descriptor languages and structural models, and complicating search. Partly this is because we can now deal

[3] Vickery claimed to have received, after the original publication of MARLIS in *American Documentation*, requests from people wanting to know how they could get hold of a MARLIS.

with entire documents already encoded in machine-readable form as full text. This in turn allows us to do full-text search without any further document analysis or description, or any control on that process – although formal descriptor languages still exist and are used in some contexts, it is entirely possible to work without them altogether. Vickery's structural models were essentially about constructing formal descriptor languages.

I would not of course claim that one *should* do without formal descriptor languages altogether, and it is very clear that some such devices (whether in the form of ontologies, knowledge bases, named entity recognisers or whatever) have important roles to play in search, particularly outside of Web-scale search, in enterprises and the like. And possibly that realisation is seeping back again into the current IR world. But the extraordinary success on the Web of methods without explicit semantic or knowledge structures has taken a large part of the IR research world along with it.

Chapter 8 of the book is 'The Automation of Storage and Retrieval', discussing the possibilities for automation and the use of machines. In the context of the time of writing, 'machine' does not necessarily mean computer, and 'automation' does not necessarily mean computer-based – indeed, computers play a relatively small role in this discussion. The main forms of automation existing at the time were based on various types of punched card, as well as ('experimentally at least') paper tape, magnetic tape, magnetic drums and disks, and photographic film. The focus on storage media is interesting – while we do not now think of Google as simply a large bank of magnetic disks, actually of course that is a very significant and substantial part of what Google is. Again in the context of the time, the medium had a very strong influence on what was possible for the representation and indexing of documents and queries and the matching process. Nevertheless, part of Vickery's purpose here is abstraction, as discussed in the next section.

The abstractions

Essentially the first part of chapter 8 is Vickery's attempt to abstract the various stages necessary in retrieval, and discuss them as abstractions rather than in terms of specific methods. Inevitably, in retrospect, these abstractions look a little strange, and very strongly embedded in the then zeitgeist, as well as in the physical nature of the systems and devices then available. Nevertheless, there is some interest in going a little deeper.

The first distinction made is between the 'original text' and the 'manipulable text' – essentially between print and machine-readable text. There is a very strong assumption that the document itself is a physical object, to be stored in a library-like store, and addressed in some way for access. Next, the manipulable text is assumed to require a sequence of transformations (reduction to 'informative statements', selection of some of these ('relevant') statements, representation in a standardised documentary language. Then these representations need to be filed in a standard way to allow for searching.

The search phase contains some of the same steps – again, search queries are seen as needing to be transformed into manipulable form. However, for the subsequent formalisation, Vickery expects that queries need human interpretation. By implication, and indeed by explicit discussion in the earlier chapter on Search Procedures, the human involved is a librarian, not the user. In the earlier chapter, Vickery quotes with approval another commentator saying 'The requester *cannot be expected* to translate his own question' (my emphasis): this describes well a basic premise that Vickery shares with many other people of the time.

The formal query then needs matching against the stored file of document representations – extensive tables discuss the interaction between the physical medium, the form of file organisation chosen, and the matching method. Then of course the results of this process are the addresses (only) of the physical documents, which need to be located and offered to the user.

Many of these steps are seen by Vickery as possibly mechanisable, though he clearly has doubts about when this will happen and how effective it will be. He does acknowledge the extent to which the analysis is based on then-current systems, commenting that his abstraction is 'essentially a mechanisation of the intellectual operations of the human indexer', and that it may be 'unnecessarily complicated'. He mentions the possibility of using methods like those developed by HP Luhn, of which more below.

Statistics

Another book I have in front of me is the proceedings of a recent conference on the *Theory of Information Retrieval* (Azzopardi et al, 2009). The concerns reflected in the papers presented at this conference have rather little overlap with Vickery's. However, it is worth mentioning one connection. One of the dominant themes of this and other contemporary IR conferences is the interest in statistical argument and models. There are now very many different approaches to IR which have substantial statistical content. Models discussed in this conference include quantum models, language modelling (that is, statistical language modelling), regression, divergence from randomness, and belief models, to name but a few; the phrase 'machine learning' is not used in any of the titles, but concepts taken from ML are pervasive. The reasons for this strong statistical influence are several. Throughout the last quarter of the twentieth century, the influence of statistical methods spread, first through the research community and then (with a bang) into public use with the development of the web search engines. As an aside, the field of linguistics, which in 1961 was dominated by formal structural approaches such as that of Chomsky, also succumbed to the spread of statistical ideas; NLP is nowhere nowadays without statistical models.

It can be argued that this process may have gone too far. Certainly the success of the web search engines is dependent on the scale on which they operate, and is hard to reproduce in the context of smaller or more specialised collections. It seems clear that statistical ideas are here to stay, and will continue to play a major role; but also that statistics is not the be-all and end-all of retrieval system theory.

So did Vickery foresee any of this? Not exactly, and it would have taken a truly remarkable prescience to have made any strong predictions on that score in 1961[4]. Nevertheless, it is possible

[4] My second edition contains a single reference to a publication in 1962 by Salton, but it is about the manipulation of tree-like thesaurus structures, nothing to do with the vector space model with which he is most closely associated. Needham's clustering work also makes a single brief appearance, but Spärck Jones is not yet in evidence.

to discover, scattered throughout the book, intimations of the possible future power of statistics. One such is the multiple references to the work of HP Luhn, mentioned above: Luhn and others were beginning to try to replace or substitute certain traditionally human/semantic/cognitive operations with basically statistical ones. Most of these attempts were largely pragmatic in nature, but Maron and others, who had just begun work on what might count as the first formal probabilistic model for IR, also make an appearance.

A mongrel field

I might be tempted to claim, given my own heavy involvement in probabilistic and statistical approaches, that any possible unified theory of IR should be statistical in nature. However, I resist this temptation. More and more, it seems to me that we should not expect IR to be reducible to one unified theory. There exists a multiplicity of theoretical domains and ways of thinking which are capable of providing valuable insights into this field of technology or engineering, and there will be more in the future. Some of these will be borrowed or purloined from other fields; some may be home-grown, or may evolve internally for re-export to other fields. This, as I see it now, is all as it should be.

I also believe that Brian Vickery would have had some sympathy with this opinion. He was nothing if not pragmatic, and was happy to temper his appeals to theoretical argument with what might now be described as reality checks.

Vickery's subsequent work

Vickery's later and much more wide-ranging work is explored elsewhere at this conference. To some extent he drew back from a focus on information retrieval in favour of a broader view, as the sequence of his later book titles would suggest: *Information Retrieval Techniques*, *Information Systems*, *Information Science in Theory and Practice*. It might also be argued that the mainstream field of information retrieval went in the opposite direction, becoming more narrowly focussed, perhaps too much so. That argument can be heard strongly today from the research community that is associated with the phrase 'information seeking'. Vickery's longstanding and consistent emphasis on users of information retrieval systems, and on keeping users in mind while discussing systems, is essentially the same argument.

References

Azzopardi, L., et al, eds. (2009). Advances in information retrieval theory. In *Proceedings of ICTIR 2009*. Springer: Berlin, 2009.

Vickery, B.C. (1965) *On retrieval system theory*. London: Butterworths, 1961. 2nd ed, 1965.

Corresponding author:

Stephen Robertson can be contacted at stephenerobertson@hotmail.co.uk