Document retrieval and text retrieval

Karen Sparck Jones

Computer Laboratory, University of Cambridge
New Museums Site, Pembroke Street, Cambridge CB2 3QG, UK

1. Essentials of document retrieval

Document retrieval (DR) is for the user who wants to find out about something by reading about it.

DR systems illustrate every variety of indexing language, request and document description, and search mechanism. Controlled languages (CLs) have been commonly used, across the range from only slightly restricted natural language (NL) to a carefully designed artificial language. With CLs professional indexing is required and professional searching is the norm. However automatic DR systems have also encouraged the use of NL through searching on titles and abstracts. This naturally makes end-user searching practicable, though not necessarily easy; and end-users often lack the experience to search effectively when strict word matching fails to identify appropriate documents. The essential requirement in retrieval is that a match between request and document descriptions should reflect the underlying relation between user need and document content.

Indexing, providing for matches, thus aims to promote precision and/or recall. It often has to do this under the external constraints of large files with small relevance sets. It always has to do it under the internal constraints on indexing itself. These are, for both requests and documents, *variability* of language, whether this stems from ambiguity or differences of perspective; for requests, *underspecification*, whether through vagueness or incompleteness; and for documents, information *reduction*, whether through generalisation or selection. Reduction is essential for DR, for both efficiency in scanning and effectiveness in concentrating on key content.

The implications of these constraints for index language design and use are conflicting, and suggest many alternative possibilities within the CL/NL space for the treatment of terms and term relations, of implicit and explicit relations, and of syntagmatic and paradigmatic relations. Mixes and tradeoffs are possible, and the necessary flexibility is achieved, because descriptions are manipulated in searching.

However though the conventional preference is for CLs, extensive tests have shown that very competitive performance can be obtained through cheap and simple indexing using coordinated single NL terms along with statistical selection and weighting, ranked output, and relevance feedback. The gains from this approach come from allowing for *late binding* and *redundancy*, along with *derivation* from source documents, in topic characterisation. The findings have been supported by many tests investigating different DR system factors, and the approach has been implemented commercially. But the test evidence is not always strong, and the tests have been on a limited scale; further, the strategy depends on request quality and probably also on working with document surrogates, like abstracts, which concentrate information. Even so, the puzzle is that linguistic sophistication, even with human LP, does not provide clear performance gains, and routine performance typically falls within an undistinguished 30-60% R-P tradeoff area.

2. Text retrieval

However a new situation has arisen with the availability of machine-readable full text. For text retrieval (TR), NLP to provide more sophisticated indexing may be needed because more discrimination within large files of long texts is required, or may be desired because more focusing is possible. This suggests the more NLP the better, but whether for better-motivated simple indexing or for more complex representation has to be determined.

Given past experience, and the need for flexibility in the face of uncertainty, a sound approach appears to be to maintain overall simplicity but to allow for more complex indexing descriptors than single terms, derived through NLP and NL-flavoured, e.g. simple phrases or predications. These would be just coordinated for descriptions but, more importantly, statistically selected and weighted. To obtain the reduced descriptions still needed to emphasise important text content, text-locational or statistical information could be exploited. To support indexing, and, more critically, searching a terminological apparatus again of a simple NL-oriented kind providing term substitutes or collocates, and again statistically controlled, could be valuable. Searching should allow the substitution or relaxation of elements and relations in complex terms, again with weighting, especially via feedback. This whole approach would emphasise the NL of the texts while recognising the statistical properties of large files and long documents. The crux is thus to demonstrate that linguistically-constrained terms are superior to e.g. co-locational ones.

Heavy testing is needed to establish performance for the suggested approach, given the many factors affecting retrieval systems, both environment variables e.g. document type, subject domain, user category, and system parameters e.g. description exhaustivity, language specificity, weighting formula. There are also different evaluation criteria, performance measures, and application methods to consider. Proper testing is hard (and costly) since it requires large collections, of requests as much as documents, with relevance assessments, and implies fine-grained comparisons within a grid of system contexts and design options.

Various approaches along the lines suggested, as well as simpler DR-derived ones, are being investigated within ARPA TREC. The TREC experiments are important as the largest retrieval tests to date, with an earnest evaluation design, as well as being TR tests on the grand scale. But any conclusions drawn from them must be treated with caution since the TREC queries are highly honed, and are for standing interests (matching a document against many requests not vice versa), with tightly specified response needs. TREC is not typical of many retrieval situations, notably the 'wants to read about' one, so any results obtained, especially good ones relying on collection tailoring, may not be generally applicable and other tests are mandatory.

3. HLT issues

In the present state of TR research, and the HLT context, the issues are as follows:

1. With respect to the *objects* manipulated in retrieval, i.e. index descriptions, given that indexing is making predictions for future searching:

   What kind of sophistication is in order: what concepts should be selected and how should they be represented? How should linguistic and statistical facts be related? For example, how should weights for compounds be derived, by wholes or from constituents, and how should matching, by wholes or constituents, be handled?

2. Wrt the *process* of retrieval, given that searching is fundamentally interactive:

   What way of developing requests is best: should the system be proactive or reactive? How can the user be involved? For example, how can the user cope with CLs that are incomprehensible (through notation) or misleading (through pseudo-English); or with statistical numbers?

3. Wrt the *implementation* of retrieval systems, given their asymmetry with requests demanding notice but many documents never wanted:

   What distribution of effort is rational: should effort be at file time or search time? How can flexibility be maintained? For instance, when should compounds be formed, or their weights computed?

4. Wrt the *model* adopted to underpin systems, given the lumpiness inherent in system operation in the mass and average but user interest in the individual and distinctive:

   What strength of assumptions is rational: should the system work with the vector, or probabilistic, or some other model? How can an abstract formal model supply specific instructions for action? For instance, can the model say precisely how matches should be scored?

5. Wrt retrieval using *full text*, given that with more detail there is also more noise:

   What functions should TR serve: should it help to refine indexing or offer passage retrieval? How might indexing and searching on two levels operate? For instance, how can a dispersed concept, spread over text, be identified?

6. Wrt system *testing*, given the enormous variety of environment factors and system possibilities:

   What degree of reality and representativeness is required for validity: can collections be picked up or must they be designed? How can control be imposed to isolate factor effects? For instance, how should non-repeatable user search data be treated?

These issues reflect the conflict between the fact of interdependencies within systems and the aim of decomposition for understanding and design. Thus the key points for DR and TR as potential NLP tasks, as opposed to e.g. database query or translation, is that *scale phenomena* count; thus the value of index descriptions is in file discrimination, not document definition; and retrieval output is contingent on the lifetime file, not the local situation. At the same time, information retrieval experience has shown that *any* approach can seem plausible, as also that whatever one does comes out grey in the wash.