Retrieval system models: what's new?

Stephen Robertson
Microsoft Research Limited, Cambridge
Karen Spärck Jones
Computer Laboratory, University of Cambridge

This paper was published in
Computer systems: theory, technology, and applications. A tribute to Roger Needham,
New York, Springer, 2004, 237-242.

Automated retrieval systems

In the postwar development of computing, most people thought of computers as machines for numerical applications. But some saw the potential for automatic text processing tasks, notably translation and document indexing and searching, even though words seemed much messier as data than numbers. For Roger, as one of these early researchers, building systems for language processing was both intellectually challenging and practically useful, and in the late 1950s he began to work on document retrieval (Needham 1963). The specialised scientific literature was growing too fast for the existing broadly-based and rigid indexing and classification schemes. This lack of appropriate retrieval tools, and the opportunities offered by computers, stimulated a critical examination of existing approaches to indexing and searching and the introduction of radically new ones.

Document (or text) retrieval systems, like libraries before them, depend on a model of the way documents should be characterised to facilitate searching, and of effective strategies for searching. Many models for retrieval systems have been proposed since the 1950s. The most innovative, attractive, and successful have been those that, unlike the earlier library models, have exploited the behaviour of the actual words used in document texts, and have facilitated flexible matching between queries and documents, leading to a ranked search output. These ground features of modern systems fit automation very well, and automation has made it possible to take advantage of the distribution of terms in documents to allow, e.g. term weighting. There are, however, different ways of modelling retrieval systems within this broad framework, and it has not been possible, until recently, to provide concrete evidence for the real value and relative merits of the competing models. It has been impracticable to conduct the necessary large-scale retrieval experiments, because performance evaluation depends on having information about which documents are relevant to a query, and getting this information is extremely expensive.

This situation has changed in a number of ways. The development of the Web and the proliferation of machine-readable text (in the broadest sense) have made the 'information layer' and its operations much more central to computing in general than they were in the 50s. 'Retrieval' is now taken to encompass a wide range of different tasks. Probably as a consequence, seriously more resources have over the last decade or two become available for

work in the general area of text retrieval. Retrieval research since Roger worked on it in the late 1950s and early 60s has changed out of all recognition.

These changes have brought the issue of models to the forefront, and have also afforded much greater opportunities for experimental work. Both these themes are explored below.

Retrieval system evaluation and model testing

The DARPA/NIST Text REtrieval Conferences (TRECs), initiated at the beginning of the 90s and still flourishing, have made it possible to evaluate retrieval systems far more thoroughly than ever before. The scale of the data in TREC, the range of tasks, the number of participants, and the multitude of tests have all contributed to this sea change.

Much of this effort has indeed gone into exploring variations on, and developments of, familiar themes, in fact ones dating back to the beginnings of automated retrieval research. But TREC has led to more than this, in two important ways. Many (though not all) of the retrieval systems tested have an explicit theoretical underpinning, or at least implicitly assume one. The Cornell Vector Space Model (VSM) is the most commonly invoked, but the University of Massachusetts Inference Model (IM), and the London/Cambridge Probabilistic Model (PM) have also been conspicuous since TREC began in 1992.

TREC has been sufficiently rigorous to subject not only system implementations based on these models, but the models themselves, to serious stress testing. The models have benefitted from the development forced on them. They have also performed very well. Newer models have appeared too. Tests with a recent and strongly-argued Non-Classical Logic Model (NCLM) have so far been limited, but Language Modelling, derived from speech recognition, has been very successfully applied in TREC to the rather different retrieval task.

All of these models operate within the generic framework mentioned in the previous section, and are statistically-based. They exploit occurrence and cooccurrence patterns in index terms and documents for term weighting, search query expansion, and the like. The fact that the models perform well, and scale up, is no longer a research surprise. Nor is the fact that they perform much the same. The basic data are all the same: there are document texts, query texts, and documents judged relevant to queries; and these are all data supplying some usable information about what retrieval is really about, namely document contents, information needs, and so forth. Further, since document retrieval is essentially an approximate task being conducted in a large and partially-understood conceptual space, the same general properties of the objects in the space matter for all the theories and invoke the same responses from all of them, as eventually reflected in $tf{*}idf$-style[1] term weighting. Several of the models also share, again not surprisingly, a generic probabilistic approach to retrieval.

But the models at their most fundamental are rather different. So we may ask how one might compare these different views, or on what grounds one might choose between them. The primary issue both of comparison and of choice is usually taken to be retrieval performance. But they may be compared in other ways, particularly in the absence of a consistent and material performance differential. We may consider the richness of each approach, in the sense of the extent to which it suggests or promotes different methods or techniques. We may, in ideal scientific fashion, attempt to make and validate experimentally further predictions from the models, other than of good retrieval performance. We may also – this is the main aim

[1] A commonly used form of term weighting which gives more importance to a term occurring frequently in the document under consideration, and less to a term which occurs in many documents

of the present note – discuss how each type of model views the critical relationships between retrieval objects (documents, queries, terms).

Model characteristics

This attempt to characterise the various models by how they see the relationship between documents and queries is of necessity crude and over-simplified, if only because it is often perfectly feasible for different theorists to accept the same formal framework on the basis of very different fundamental assumptions or interpretations. However, what follows may be a useful sketch.

The VSM treats the query-document relationship simply as an object *proximity* relation in an information space. There may be other objects associated with the space, like index terms. The vectors characterising objects (or the dimensions of the space itself, as in Latent Semantic Indexing) are manipulated to bring queries and relevant documents closer together (Salton et al. 1975).

The IM views the query-document relationship as a *connectivity* one. The connections that can be made between the two, e.g. through terms, justify the inference that a document should be retrieved (Turtle and Croft 1990).

The NCLM takes the query document relationship as a *proof* one, with the document proving the query, e.g. through statements about their index term descriptions (van Rijsbergen 1986).

The PM has a *generative* relation from a query to a document, making a prediction that a document, e.g. because it has certain terms, belongs to the class of relevant documents (Robertson et al. 1981).

In the LM there is also a *generative* relationship, but the other way round, from the document to the query, i.e. the query is thought of as derived from the document in the same sort of way that in speech the heard sounds are generated from a word string (Berger and Lafferty 1999, Miller et al. 1999).

From these broad descriptions, it may not be clear whether or not the differences are fundamental, or how important they are practically speaking. The comparison may be further confused by other similarities between them, for instance because in the IM inference is probabilistic, or because the PM may be given a network implementation (Kwok 1995). One difference which does appear fundamental lies in whether the key retrieval notion of relevance figures explicitly as a model primitive. It does this in the PM, so that the generation relation is actually from both query *and* relevance to a retrieval-worthy document. Relevance does not figure so explicitly in the VSM, or in the IM or NCLM. We have argued elsewhere (in Croft and Lafferty in press) that the LM does not explicitly use relevance either (although it has more recently been presented with an explicit relevance variable included in the model - see Lafferty and Zhai in Croft and Lafferty).

But though relevance may be taken as a primitive in a model, strictly relevance is inaccessible, a hidden variable, and at a very practical level, all the models may be interpreted as saying that the stronger the proximity/connectivity... relation between query and document is, and thus the more highly ranked a document is in the search output, the more likely it is that the user will find the document relevant to their information need. Furthermore, for all the models, the specific expression of this proximity/... notion always makes use of the same basic statistical facts.

Model implications

The point just made does not, however, imply that the models are mere notational variants of one another. They indeed all deal in the same objects, queries, documents, terms etc, and all (in one way or another, and in various versions) respond to the statistical properties of retrieval data. But they make use of notions that are individually distinctive, albeit very general. So one question is whether any of the ground notions like proximity, inference, generation, etc is more intuitively satisfying as a (or perhaps the) key concept for a theory of retrieval. Such a question may be taken as essentially a metaphysical matter, but another question is whether thinking about retrieval systems in terms of one central notion rather than another is more productive as a base for building effective (and robust etc) systems.

One possible position here is that the fact that some generic model has been used for different information and language processing tasks is important, because it reflects the fact that these tasks are all, broadly speaking, discourse (text) transformation tasks with something in common. From this point of view LM, which has been applied to translation and summarising as well as speech transcription and retrieval, has something going for it. But on inspection, the LM generative account for some of these tasks seems distinctly forced. Other model mechanisms, like vector operations or the use of Bayes' Theorem, have been very widely exploited, but these are too abstract to make substantive task links in the way that language modelling is claimed to do through the idea of generation.

However another view is that even if there are genuine differences between the abstract models, this doesn't really matter because it is not where the shoe pinches. Thus consider the three input contributors to a retrieval system: the formal model (F); the estimation accuracy (or training potential) of the model (E); and the implementation detail (I). As already noted, when it comes to I, the weighting formulae used, for example, are much the same. With F, on the other hand, there either are no real differences, or the only differences that count are those that affect E, since this is what is going to determine operational system effectiveness. Any system using any model, in the statistical retrieval world, has to exploit its known data to predict what documents will be valuable. It may be that the LM approach (with a variety of different applications already developed) has an advantage here, in the form of a rich range of estimation methods on which to draw.

With the evaluation data we now have, we are in a much better position to assess claims of this kind. We can hope to demonstrate whether any of the models is superior to the others, either because its key notions are more productive in leading to good ways of looking at different retrieval tasks, or because it provides better ways of dealing with the challenges of estimation, or even because it leads to better performing implementations in, say, choice of weighting formulae. The question of what a retrieval system should be like, in its essentials, was one that Roger worked on, and his work was one of the sources of a modern probabilistic system (Sparck Jones et al. 2000). Just as we benefited from his comments in the past, so we would have welcomed his views on the present Retrieval Model Action Space.

References

Berger, A. and Lafferty, J. 'Information retrieval as statistical translation', *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, 222-229.

Croft, W.B. and Lafferty, J. (Eds.) *Language modelling for information retrieval*, Dordrecht: Kluwer, in press.

Kwok, K.L. 'A network approach to probabilistic information retrieval', *ACM Transactions on Information Systems*, 13, 1995, 324-353.

Miller, D.R.H. Leek, T. and Schwartz, R.M. 'A hidden Markov model retrieval system', *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, 214-221.

Needham, R.M. 'A method for using computers in information classification', *Information Processing 62: Proceedings of IFIP Congress 1962*, (Ed. Popplewell), Amsterdam: North-Holland, 1963, 284-287.

Rijsbergen, C.J. van 'A non-classical logic for information retrieval', *The Computer Journal*, 29, 1986, 481-485.

Robertson, S.E., van Rijsbergen, C.J. and Porter, M.F. 'Probabilistic models of indexing and searching', in *Information retrieval research*, (Ed. R.N. Oddy et al.), London: Butterworths, 1981, 35-56.

Salton, G., Wong, A. and Yang, C.S. 'A vector space model for automatic indexing', *Communications of the ACM*, 18, 1975, 613-620.

Sparck Jones, K., Walker, S. and Robertson, S.E. 'A probabilistic model of information retrieval: development and comparative experiments. Parts 1 and 2', *Information Processing and Management*, 36, 2000, 779-840.

Turtle, H.R. and Croft, W.B. 'Inference networks for document retrieval', *Proceedings of the13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1990, 1-24.