IDF term weighting and IR research lessons

Karen Sparck Jones
Computer Laboratory, University of Cambridge

This paper appeared in *Journal of Documentation*, 60, 2004, 521-523.

Abstract

Robertson comments on the theoretical status of IDF term weighting. Its history illustrates how ideas develop in a specific research context, in theory/experiment interaction, and in operational practice.

It is an honour to have the small proposal for term weighting that I published more than thirty years ago (Sparck Jones 1972) the subject of Stephen Robertson's paper (Robertson 2004). I would like to comment on some points that I see as suggesting lessons for information retrieval research.

First, the context that prompted the proposal.

The proposal came from trying to explain why earlier ideas about how to do automatic indexing did not work. They were plausible in themselves, but had quite different objectives. My previous research had concentrated on automatic methods for constructing term classifications intended, by analogy with manual thesauri, as recall-promoting devices. Classes were based on term cooccurrences in documents, following the generic statistical approach to retrieval initially suggested by Luhn, and applied within the coordination level matching framework. But these classifications, on Cleverdon's Cranfield data and using the test and evaluation methods that he and Salton had established, did not deliver the predicted improvements in retrieval performance. The best performance was obtained with (necessarily) small groups of very similar terms.

Trying to understand what was happening in detail showed that terms that occurred in many documents dominated the classes. Thus anything that increased their matching potential, as term substitution did, would inevitably retrieve non-relevant rather more than relevant documents. However these frequent terms were also common in requests, and simply removing them, as advocated by Svenonius, could have a damaging effect on performance. The natural implication was therefore that less frequent terms should be grouped but more frequent ones should be confined to singleton classes. This could give better performance than terms alone, but not for all test collections.

What all this suggested was that it might be more profitable to concentrate on the frequency behaviour of terms, and forget about classes. More specifically, it led to the idea that all terms should be allowed to match but the value of matches on frequent terms should be lower than that for non-frequent terms. Roger Needham, who had earlier worked on statistically-based methods of indexing and retrieval, was easily able as a mathematician to suggest an appropriate simple formula that smoothly damped down frequency and was shown to work reliably and usefully for different collections.

My proposal for weighting was thus a direct response to the results of the kind of systematic retrieval testing that Cleverdon did so much to establish. It was also, when compared with Salton's work on automatic indexing, a product of subtly different data conditions. Salton

worked on indexing for abstracts or short full texts, and thus early saw that within-document term frequency, *tf*, could be important. My input was simple presence/absence manual word lists, which naturally led me to consider the effects that the number of documents a term occurred in could have.

Second, the associated theory.

As Stephen's paper says, the *idf* proposal was not set within any specific theoretical framework, or at least anything stronger than the general idea that term distribution patterns could be correlated with retrieval value, that dated back to Luhn. In the earlier 1970s Salton and his students explored the generic statistical approach within the overall vector model of information space, and elaborated the idea of term discrimination value, which distinguished high, medium and low frequency value rather than just higher and lower. Salton also allowed for document (and query) length but its implications were not really recognised, since there were no serious full-text test collections. Properly incorporating *dl* in weighting formulae had to wait for the kick delivered later by the TREC test conditions.

It was, however, apparent that with relevance information, facts about data distributions of the kind that that motivated *idf* could be made much more effective. Salton's group had already explored relevance feedback, but in a very different style. In my case, experiments initially motivated by Miller's work stimulated the productive interaction between testing and theory development that led to Robertson and Sparck Jones (1976). This research in turn encouraged the subsequent work on the probabilistic retrieval model that has both given a formal context for *idf* and, particularly under TREC test pressure, has extended and consolidated the model, as Stephen's paper describes (as it also shows how tricky it is to get the theory right). Other important retrieval models do not necessarily use *tf*idf*-type weighting explicitly, but they respond to the data verities that underlie it in an analogous way.

Third, when an idea's time comes.

As is well known, operational bibliographic services were very reluctant to allow statistical methods any possible utility, especially given the tiny research experiments, and became substantially committed to the conventional boolean approach. The first Web engine builders had no such prior commitments and picked up the statistical idea: thus AltaVista applied *tf*idf* from the start, and it seems that most engines, somewhere, use something of the sort as one component of their matching strategies. It thus took about twenty five years for a simple, obvious, useful idea to reach the real world, even the fast-moving information technology one. At the same time, there has been a striking new development in the last decade, stimulated by TREC and related language and information processing evaluations that have explored a range of tasks including topic tracking, question answering, and summarising. These tasks rely, in various ways, on identifying key texts or text segments for more detailed attention, and *tf*idf* has proved a very handy tool for this purpose. The statistical successes of the retrieval world, and the way text data figure ever more largely in both system training and system application, have together propagated *tf*idf*, like a benign virus, to many places that even visionary believers in automated information processing in 1972 would never have envisaged.

References

Cleverdon, C.W. 'The Cranfield tests on index language devices,' *Aslib Proceedings*, 19, 1967, 173-194.

Luhn, H.P.: see Schultz, C.K. (ed.) *H.P. Luhn: Pioneer of information science*, New York: Spartan Books, 1968.

Miller, W.L. 'Probabilistic search strategy for Medlars', *Journal of Documentation*, 27, 1971, 254-266.

Robertson, S.E. 'On theoretical arguments for IDF', *Journal of Documentation*, 2004.

Robertson, S.E. and Sparck Jones, K. 'Relevance weighting of search terms', *Journal of the American Society for Information Science*, 27, 1976, 129-146.

Salton, G. *Automatic information organisation and retrieval*, New York: McGraw-Hill, 1968.

Salton, G. *A theory of indexing*, Philadelphia, Society for Industrial and Applied Mathematics, 1975.

Sparck Jones, K. 'A statistical interpretation of term specificity and its application in retrieval', *Journal of Documentation*, 28, 1972, 11-21.

Svenonius, E. 'An experiment in index term frequency,' *Journal of the American Society for Information Science*, 23, 1972, 109-121.

TREC: see Voorhees, E.M. and Harman, D.K. (eds.) *TREC: experiment and evaluation in information retrieval*, Cambridge, MA: MIT Press, in press.