Revisiting classification for retrieval

Karen Spärck Jones
Computer Laboratory, University of Cambridge

This paper appeared in *Journal of Documentation*, 61, 2005, 298-601

Abstract

Hjørland and Pedersen argue for a pragmatic approach to classification for retrieval, that for them implies recognition for human expertise. But a text-based approach to classification may better reflect actual conceptual structures in a field, provided that any automatic techniques used are well-motivated and formally sound.

I was interested to read the argument that Hjørland and Pedersen develop in (1) from the starting point of my 1970 paper (2), exploiting two orthogonal polarities - positivist vs. pragmatic, and theoretical vs. practical. I would like to comment briefly from two points of view: first, in relation to the larger context within which the original paper was written and second, in relation to developments in natural language information processing (NLIP) in the last decade.

The original context

Hjørland and Pedersen maintain that I oscillate between positivism and pragmatism in (2). I would argue that my emphasis was more pragmatic than positivist. The purpose for which a classification is required is what really counts. But I wanted to draw attention to the other side of things, the need to have soundly-based abstract forms of classification, and the need to understand these so as to be able to make appropriate choices for application purposes.

Being able to do anything in a fully formal and totally geared way, working from retrieval purpose to classification choice, is an ideal that is only attainable in constrained circumstances. But thinking about how it might be done is a valuable discipline. My paper was written at a time when automatic classification techniques had been proliferating and with them the danger, for those with application needs, of what might be called the flower stall temptation: I'll take a bunch of those today. It was essential to look behind plausible algorithms, to examine the generic forms of classification they involved and their methodological grounding.

My concern was essentially for more analytic care about both elements of classification for retrieval: the retrieval task and the classification strategy. I approached it by inviting more thought about the classification omponent because that appeared to be the more pressing point at the time. I took much about the nature of the retrieval task for granted, when indeed much of this was also shared between those working on manual or on automatic classification for. However my experience not long after compelled me to pay more attention, myself, to the task assumption on which my work on automatic classification for retrieval was based. I was driven, by disappointing performance results, to reexamine the notion that classification was to promote recall, and to look more carefully at how the properties of the data affected the classes obtained (see e.g. (3)), even the potential for classification (see (4)).Thus referring in (2) to the abstract requirement that classification should minimise distortion has to be taken

together with a proper assessment of the basic data description to which it is applied, though I did not consider this issue there.

Overall, though I emphasised formal theory as a desideratum in (2), the discussion was primarily about theory in a broader sense, as required to motivate formal theory. Moreover, though the focus was on the classification side, there was a theoretical underpinning on the retrieval side. This theoretical underpinning had emerged during the sixties under the stimulus of automation, but with its own independent justification (and indeed Wittgensteinian imprimatur), and some moves had been made to formalise it as needed to build the bridge with classification methods. In broad terms, this theory was that meanings and concepts are emergent from the way words are used in discourse. It is thus relevant to Hjørland and Pedersen's own views about concept analysis and understanding (and eventually classification) for IR, and also relates to developments in NLIP in the last decade.

The current context

The world is now awash with machine-readable text. This provides far more evidence about how words are used, over time, than was ever available before and is being increasingly exploited, applying the same generic principles (what you say is what you mean) for a range of NLIP tasks that have much more connection, in execution as well as perception, than before. Thus text retrieval, fact extraction, question answering and summarising share notions and processes. IR is not in a box of its own.

Hjørland and Pedersen argue for a substantive theory of classification for IR, i.e. a broad theory with its own criteria for what classification for retrieval is about, drawing on the example of medicine. In the early work on automatic classification this issue also arose, bearing particularly on the choice of basic descriptive features - even in the biological case it was accepted that recovering phylogeny need not be the primary goal. All of the early work on automatic classification recognised that the initial choices of objects, and more especially properties, to use were crucial and pragmatically loaded. Recognising the role of the contexts in which classifications are built, and the purposes for which they are built, does not, however, imply that it is necessary to invoke human experts for an initial, deep 'pre'- or 'proto'-classification analysis of literatures themselves to ensure correct automatic classification. My original paper only argued for choices of model informed by classificatory purpose and, naturally, some understanding of the generic character of the material being classified. This is rather different from the idea that automatic classifications should be built to conform to some idea of what the classifications should descriptively capture.

In particular, the argument against letting humans instruct actual classifications is strengthened by recent NLIP experience. It is impossible for humans to avoid hidden biases, i.e. not all attitudes grounded in more knowledge than a program has, but those that are unrecognised and, when inspected, quite possibly ill-grounded. These are especially dangerous when they apply to classifications that are to be used *predictively*, as retrieval classifications, and the indexing that they support, are used. As researchers in NLIP have found, exploring the ways words are actually used in large text corpora shows that they reflect different, and changing, paradigms or cultural worlds. Classifications based on massive text data can in principle be better mirrors of what is going on, in some one or several worlds, than even a well-trained individual may be able to establish when dealing with really large data, and may therefore be more reliable for predictive application. This text-based approach can of course extend in principle to exploiting more purpose-oriented discourse characteristics (e.g. as embodied

in genre), even if these have not typically been extensively investigated in automatic classifications for retrieval. Thus one can make a good case for a pragmatic view that starting with rather minimal assumptions about objects and properties, along with an approach to classification form motivated by some generally-stated classificatory purpose, and seeing how users respond to what a retrieval system using some classification delivers, is a better strategy than trying to figure out in advance, on an unavoidably more intuitive basis, what such a classification ought to be like.

This is not to deny that there are many particular information management situations where the implicit has to be made explicit, or where a richer descriptive apparatus is required, as in the special technical contexts for which ontologies are currently being advocated, for which a human contribution is essential. I also accept that many of the functional features to which Hjørland and Pedersen refer as bearing on useful classification, while accessible in principle to derivation from massive raw data, may be less so in practice from the actual available data, and thus need to be supplemented. But recent research in machine learning applied to Web data shows it can recover surprisingly refined information. Even here, however, Hjørland and Pedersen and I agree on the fact that the choice of specific classification method has to be grounded not merely in an understanding of its generic properties, but on establishing a proper link with the motivating purpose and with some broad theory about what that purpose involves. The latter has also to justify, in the finer grain and in a way that is at least gestures towards the formal, how the actual classification method used is operationalised.

All of this is hard. The classical line of IR experiment has not dealt with many of the issues that operational practice has to tackle. But the interaction between theory and experiment that major NLIP evaluation programmes, notably the Text REtrieval Conferences (5), have encouraged, have significantly advanced our understanding of IR. The text-based approach to IR is of course also rampant in the new practical reality of the Web engines.

At the same time, one of the most important techniques developed in retrieval research and very prominent in recent work, namely relevance feedback, raises a more fundamental question. This is whether classification in the conventional, explicit sense, is really needed for retrieval in many, or most, cases, or whether classification in the general (i.e. default) retrieval context has a quite other interpretation. Relevance feedback simply exploits term distribution information along with relevance judgements on viewed documents in order to modify queries. In doing this it is forming and using an implicit term classification for a particular user situation. As classification the process is indirect and minimal. It indeed depends on what properties are chosen as the basic data features, e.g. simple terms and, through weighting, on the values they can take; but beyond that it assumes very little from the point of view of classification. It is possible to argue that for at least the core retrieval requirement, giving a user more of what they like, it is fine. Yet is certainly not a big deal as classification per se: in fact most of the mileage comes from weighting. And how large that mileage can be is what retrieval research in the many experiments done in the last decade have demonstrated, and Web engines have taken on board.

(1) Hjørland, B. and Pedersen, K.N. 'A substantive theory of classification for information retrieval', *Journal of Documentation*, XXX.

(2) Spärck Jones, K. 'Some thoughts on classification for retrieval', *Journal of Documentation*, 26, 1970, 89-101.

(3) Spärck Jones, K. and Barber, E.O. 'What makes an automatic classification effective?', *Journal of the American Society for Information Science*, 22, 1971, 166-175.

(4) van Rijsbergen, C.J. and Spärck Jones, K.
"A test for the separation of relevant and non-relevant documents in experimental retrieval collections", *Journal of Documentation*, 29, 1973, 251-257.

(5) Voorhees, E.M. and Harman, D.K. (Eds.) *TREC: experiment and evaluation in information retrieval*, Cambridge MA: MIT Press, 2005.