A look back and a look forward

Karen Sparck Jones
Computer Laboratory, University of Cambridge

This paper was given for the SIGIR Award (now the Salton Award) at SIGIR 1988;
the final version appeared in
Proceedings of the 11th International Conference on Research and Development in Information Retrieval, ACM-SIGIR, 1988, 13-29.

This paper is in two parts, following the suggestion that I first comment on my own past experience in information retrieval, and then present my views on the present and future.

Some personal history

I began serious work in IR in the mid sixties through one of those funding accidents that afflict everyone in research; but I had become involved with it before that, for respectable intellectual reasons. The group working under Margaret Masterman at the Cambridge Language Research Unit had argued for the use of a thesaurus as a semantic interlingua in machine translation, and had then seen that a thesaurus could be used in a similar way, as a normalising device, for document indexing and retrieval (Masterman et al 1958). My doctoral research was concerned with automatic methods of constructing thesauri for language interpretation and generation in tasks like machine translation; and Roger Needham was working at the same time on text-based methods of constructing retrieval thesauri, in the context of research on general-purpose automatic classification techniques.

The essential common idea underlying this work was that word classes, defining lexical substitutibility, could be derived by applying formal clustering methods to word occurrence, and hence cooccurrence, data (Sparck Jones 1971b). In the early sixties we saw semantic interlinguas, thesauri, and statistical classification as promising new forms of older ideas which were well suited to the challenges and the opportunities computers offered both for carrying out language-based information managament, as in translation or retrieval, and for providing the tools, like thesauri, needed for these information extraction and transformation processes.

In my doctoral research (Sparck Jones 1964/1986) I suggested that a thesaurus could be built up by starting from sets of synonymous word senses defined by substitution in sentential text contexts, and carried out classification experiments to derive larger groups of related word senses constituting thesaurus classes from these, though I was not able to test any of my classifications as a vehicle for their ultimate purpose, namely translation. In my first major project in IR I also worked on automatic thesaurus construction, but in this case with word classes defined not through direct substitution in restricted sentential contexts, but by cooccurrence in whole texts. This rather coarse-grained classification, of the type originally studied by Roger Needham, seemed to be appropriate for document indexing and retrieval purposes. Substitution classes not confined to synonyms, but extending to collocationally related items, could be used as indexing labels within the coordination matching framework

that I have always thought natural for derivative indexing. Word classes based on text cooccurrence naturally pick up collocationally linked pairs, and capture synonym pairs only via their common collocates, but we argued that substituting a collocate is legitimate and indeed that to respond effectively to the very heterogeneous ways a concept can be expressed in text, it is necessary to allow for very heterogeneous word classes.

But it soon became clear that plausible arguments are not enough in IR. The project we began in 1965 was designed to evaluate automatic classification not only in the sense of demonstrating that classification on a realistic scale was feasible, but of showing that it had the intended recall effect in retrieval. We were therefore working with the Cranfield 2 material and began to do experiments with the smaller Cranfield collection, constructing term classifications and testing them in searching. At the CLRU we had always emphasised the need for testing in the language processing and translation work; and in the classification research, because this was concerned with automatic methods, there was a similar emphasis on testing. The importance of IR in this context was not only that it supplied challenging volumes of data, but that it came with an objective evaluation criterion: does classification promote the retrieval of relevant documents? Performance evaluation for many language processing tasks is an intrinsically difficult notion (1986a), and natural language processing research in general had in any case not advanced enough to support more than very partial or informal evaluation; while with many other applications there are no good, independent evaluation criteria because classification does not have the well-defined functional role it does in retrieval.

In earlier research on classification methods Roger Needham had already stressed that equally plausible arguments could be advanced for very different forms of classification, and we found the same for the specific IR application. More generally we found that things did not work out as we expected, and we found it very difficult to see why. The major evaluation work of the sixties, like the Cranfield and Case Western investigations and Salton's comparative experiments, showed how many environmental or data variables, and system parameters there are in an indexing and retrieval system. But we found that in trying to understand what was happening in our classification experiments, and to design experiments which would be both sufficiently informative about system behaviour and well-founded tests for particular techniques, we were driven to a finer descriptive and analytic framework which made the whole business of experiment very demanding. The same trend is clear in the Cornell research. The attempt to identify all the relevant variables and parameters, even within the relatively restricted indexing and searching area of IR systems as wholes within which we worked, that is to find an appropriate granularity in describing system properties, was a long haul driven by the need to understand system behaviour sufficiently to provide the controls required for automatic processes which have to be fully and precisely specified.

In the late sixties we concentrated on those variables and parameters most obviously relevant to automatic classification, namely the distributional properties of the term vocabulary being indexed, and the definitional properties of the classification In earlier reports I referred to environmental parameters and system variables: I think my present usage is preferable. techniques being applied, in the attempt to get an automatic classification which worked. I succeeded in this (Sparck Jones and Jackson 1970, Sparck Jones 1971a) and was able to obtain decent performance improvements with automatic classifications meeting certain requirements, restricting classification to non- frequent terms and classes to very strongly connected terms; and these results could be explained in terms of the way they limited the new terms entering document and request descriptions to ones with a high similarity in potential

relevant document incidence to the given terms.

However subsequent very detailed analytic experiments (Sparck Jones and Barber 1971) designed to discover exactly what happened when a classification was used and hence what the optimal classification strategy was, added to the earlier experience of not being led astray by plausible arguments for specific forms of classification by suggesting that the general argument for keyword clustering as a recall device might be suspect. Thus is appeared that a term classification could usefully function as a precision device.

But good-looking results for one collection were clearly not enough. We were interested in generally applicable classification techniques and, further, in classification with an operational rather than a descriptive role. So, following the tradition established at Cornell, I began comparative tests with other collections.

This led to a very complex period of research, because I found that classification was less effective on these other collections than it had been for the Cranfield one, but it was very difficult to find out why. I wanted to show that a keyword classification, constrained and applied as in the Cranfield case, would help performance. The fact that it did not provoked a long series of analytic experiments designed to uncover the influences on classification behaviour, taking the characterisation of collections and devices down to whatever level of detail seemed to be required to support the specification of effective strategies (e.g. Sparck Jones 1973a).

One outcome of this research was the Cluster Hypothesis Test (van Rijsbergen and Sparck Jones 1973). It turned out in some cases to be so difficult to get any kind of performance improvement over the term matching baseline as to suggest that it was not the devices being applied but the collection to which they were being applied that was intrinsically unrewarding.

But the main results of this work of the early seventies were those concerned with index term weighting. The research on classification led us to take an interest in the distributional properties of terms, partly for their possible effects on classification (so, for example, one shouldn't group frequent terms), and partly because term matching without the use of a classification provided a baseline standard of retrieval performance; and we found that collection frequency weighting (otherwise known as inverse document frequency weighting) was useful: it was cheap and effective, and applicable to different A program bug meant the specific results reported here were incorrect: see Sparck Jones and Bates 1977b; but the corrected results were very similar, and the test remains sound. collections (Sparck Jones 1971c, 1973b).

I nevertheless felt that all these various findings needed pulling together, and I therefore embarked on a major series of comparative experiments using a number of collections, including one large one. I still did not understand what was happening in indexing and retrieval sufficiently well, and thought that more systematic comparative information would help here: it could at least show what affected performance if not explain why or how. I also wanted to be able to demonstrate that any putative generally applicable techniques were really so. Moreover for both purposes, I wanted to feel satisfied that the tests were valid, in being properly controlled and with performance properly measured. I believed that the standard of my own experiments, as well as those of others, needed to be raised, in particular in terms of collection size, both because small scale tests were unlikely to be statistically valid and because, even if they were, the results obtained were not representative of the absolute levels of performance characteristic of large collections in actual use.

The effort involved in these tests, the work of setting up the collections and the persistent obstacles in the way of detailed comparisons with the results obtained elsewhere, were all begetters of the idea of the of Ideal Test Collection (Sparck Jones and van Rijsbergen 1976, Sparck Jones and Bates 1977a) as a well-founded community resource supporting at once

individually satisfying and connectible experiments.

The major series of tests concluded in 1976 (Sparck Jones and Bates 1977b) covered four input factors, four indexing factors and three output factors each, and particularly the indexing factors, covering a range of alternatives; fourteen test collections representing different forms of primary indexing for four document and request sets; and nine performance measurement procedures: there were hundreds of runs each matching a request set against a document set. I felt that these tests, though far from perfect, represented a significant advance in setting and maintaining experimental standards. I found the results saddening from one point of view, but exciting from another. It was depressing that, after ten years' effort, we had not been able to get anything from classification. But the line of work we began on term weighting was very interesting. Collection frequency weighting was established as useful and reliable. This exploited only the distribution of terms in documents, but Miller and subsequently Robertson had suggested that it was worth looking at the more discriminating relative distribution of terms in relevant and non-relevant documents, and this led to a most exhilarating period of research interacting with Stephen Robertson in developing and testing relevance weighting (Robertson and Sparck Jones 1976). The work was particularly satisfying because it was clear that experiments could be done to test the theory and because the test results in turn stimulated more thorough theoretical analysis and a better formulation of the theory. The research with relevance weighting was also worthwhile because it provided both a realistic measure of optimal performance and a device, relevance feedback, for improving actual performance.

The results we obtained with predictive relevance weights were both much better than those given by simple terms and much better than we obtained with other devices. My next series of experiments was therefore a major one designed to evaluate relevance weighting in a wide range of conditions, and in particular for large test collections, and to measure performance with a wide variety of methods. This was a most gruelling business, but I was determined to reach a proper standard, and to ensure that any claims that might be made for relevance weighting were legitimate. These tests, like the previous ones, involved large numbers of variables and parameters; and they, like the previous ones, required very large amounts of preliminary data processing, to derive standard-form test collections from the raw data from various sources, for example ones representing abstracts or titles, or using regular requests or Boolean SDI profiles; setting up the subsets for predictive relevance weighting was also a significant effeort. The tests again involved hundreds of runs, on seven test collections derived from four document sets, two of 11500 and 27000 documents respectively, with seven performance measures.

But all this effort was worthwhile because the tests did establish the value of relevance weighting, even where little relevance information was available Sparck Jones 1979a, Sparck Jones and Webster 1980). It was also encouraging to feel that the results had a good theoretical base, which also applied to the earlier document frequency weighting, and which was being further studied and integrated into a broader probabilistic theory of indexing and retrieval by my colleagues Stephen Robertson and Keith van Rijsbergen and others.

I felt, however, somewhat flattened by the continuous experimental grind in which we had been engaged. More importantly, I felt that the required next step in this line of work was to carry out real, rather than simulated, interactive searching, to investigate the behaviour of relevance weighting under the constraints imposed by real users, who might not be willing to look at enough documents to provide useful feedback information. Though we had already done some laboratory tests designed to see how well relevance weighting performed given little

relevance information (Sparck Jones 1979b), something much nearer real feedback conditions was required. I hoped, indeed, that the results we had obtained would be sufficiently convincing to attract those engaged with operational services, though implementing relevance weighting in these contexts presents many practical difficulties.

I was at the same time somewhat discouraged by the general lack of snap, crackle and pop evident in IR research by the end of the seventies, which did not offer stimulating new lines of work. I had maintained my interest in natural language processing, and this was manifestly then a much more dynamic area. I therefore returned to it, through a project on a natural language front end for conventional databases, though I maintained a connection with IR through the idea of an integrated inquiry system described in the second part of this paper. I further became involved with the problems of user modelling (Sparck Jones 1987) which, in its many aspects and as a general issue in discourse and dialogue processing, has become an active area of language processing research. This has also been recognised as a topic of concern for IR, which provides an interesting study context for work on the problems involved and for research on the related issues of interface architectures, that I shall consider further in the second part of this paper.

I think it a fair judgement, in reviewing all the research I have described, to say that it did show that distributional information could be successfully exploited in indexing and searching devices, and that it helped to establish experimental standards. But throughout I owed a great deal to the examples set by Cyril Cleverdon, Mike Keen and Gerry Salton, and to the productive exchanges and collaborations I have had with them and with other close colleagues, notably Keith van Rijsbergen and Stephen Robertson, as well to my research assistants of the seventies, Graham Bates and Chris Webster.

1 Thoughts on the present and future

The work I have described directly reflects the dominant preoccupations of research on automatic indexing and retrieval from the time in the late fifties when computers appeared to offer new possibilities in the way of power and objectivity. It was concentrated on the derivation of document and request descriptions from given text sources, and on the way these could be manipulated; and it sought to ground these processes in a formal theory of description and matching.

But these concerns, though worthy, had unfortunate consquences. One was that, in spite of references to environmental parameters and so forth, it tested information systems in an abstract, reductionist way which was not only felt to be disagreeably arid but was judged to neglect not only important operational matters but, more importantly, much of the vital business of establishing the user's need. Relevance feedback, and a general concentration on requests rather than documents as more worthy of attention in improving performance (following the Case Western findings of the sixties) went some way towards the user, but did nothing like enough compared with the rich interaction observed between the human intermediary and the user. The neglect of the user does not invalidate what was done, but it suggests it plays a less important part in the information management activity involved in running and using a body of documents than the concentration on it implied. The rather narrow view was however also a natural consequence of the desperate struggle to achieve experimental control which was a very proper concern and which remains a serious problem for IR research, and particularly the work on interactive searching to which I shall return

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.