SUMMARY

I Main Aims and Findings

The main object of the project was to carry out more systematic comparative tests of the automatic indexing techniques previously studied, including tests on a significantly larger scale than earlier ones. It was also intended that the work should be more directly related to that of the SMART Project, which has been concerned with automatic indexing for more than a decade.

The project has thus been concerned with the statistical techniques which have been the obvious means of achieving automatic indexing. These techniques exploit information about the occurrence and cooccurrence of potential and actual indexing keys, and hence cover term selection and weighting on the one hand, and grouping on the other. The research has also been concerned with the characteristics of the material input to the statistical processing, i.e. with the effect of the source from which index terms are derived, and the amount of information initially supplied for each document, on weighting and classification; and with the way in which indexing involving weighting and classification is utilised in searching.

The research therefore follows on from the earlier work on weighting described in Sparck Jones 1972 and 1973c, that on classification reported in Sparck Jones 1970 and 1971a and b, and that on collection characteristics and description exhaustivity of Sparck Jones 1973a and b. Some of the results presented in these publications are included in this report, for reference.

The project itself has involved experiments with a range of different test collections, including a large one consisting of 182 requests and 27361 documents, obtained from UKCIS. Several of these collections have been used by other projects, so comparisons have been made with their findings.

We have in fact found that weighting techniques are more effective than classification ones, and the main project effort has gone into developing and testing these. In particular the suggestion that relevance information can be incorporated into weighting schemes has been followed up in a series of experiments. The development of the idea, and initial tests, are reported in Sparck Jones 1975 and Robertson 1976. The approach is of interest not only because it leads to substantial performance improvements, particularly with the kind of search techniques generally used in information retrieval experiments, but because it has a sound theoretical basis. Work on somewhat similar lines has recently been carried out by Salton and the SMART Project workers (Salton 1976) and the approach is currently being taken further by van Rijsbergen (van Rijsbergen 1977). Tests applying our techniques to different data and in different retrieval contexts are described in the Report, and the results are related to SMART Project findings. Specifically, our tests show that

quite limited amounts of information about the occurrences of terms in documents and in relevant documents can be effectively used to improve performance, i.e. can be used in a predictive way to improve performance, measured by recall and precision, by 100 - 500 percent. The techniques appear to be of no value only when requests have been extremely carefully prepared and elaborately formulated.

Term classification has not proved helpful in larger experiments, for reasons which are not wholly obvious, and the project has therefore been less concerned with classification than was intended. It is discussed in the Report mainly for completeness.

The specific results presented in the body of the Report will hopefully permit a better assessment of the real potential of non-trivial automatic indexing methods than has hitherto been possible. However, we have been disappointed to find that the conclusions which can be drawn from the tests regarded as most important, namely those with a really large data set, are limited by a property of the collection which was not manifest when we took over the raw material. This is the very restricted relevance information available: it is confined to assessments of the comparatively small outputs of searches based on refined specifications. It is therefore impossible to determine performance for indexing or searching techniques materially different from those used to generate the original data, since these typically produce very different output.

A second difficulty in assessing the value of the project results as a whole as evidence for automatic indexing, is that operational retrieval systems have been rapidly developing in a direction, namely that represented by on-line searching, which could not, for good practical reasons, be taken into account in our experiments: local resources do not permit significant interactive computing, and simulation on a large scale would be very expensive; there are also problems about conducting controlled comparisons between search procedures when assessments are made during searching. As a result it is not clear how far our conclusions on automatic indexing and searching would hold for this type of retrieval system. Our findings on term weighting incorporating relevance information nevertheless appear to be sufficiently solid for us to take an optimistic view of their potential, particularly since its application to such systems is quite straightforward.

The work described in the Report may thus be summarised under lists of aims and findings, as follows.

Aims

1. To test automatic indexing techniques involving statistical weighting and/or classification in systematic comparisons between different collections of documents and requests, and on a reasonably large scale.

2. To consider the effect on the performance of these techniques of properties of the input material such as indexing source and exhaustivity.

3. To consider the effects on their performance of different methods of exploiting the information supplied by these statistical techniques in searching.

4. To correlate the project experiments and results with those of others.

Findings

1. Statistical weighting is effective, leading to noticeable and even material improvements in performance; but classification is not generally effective.

2. Variations in input data properties are not typically major influences on performance.

3. Searching techniques suited to the use of statistically-derived information and leading to an ordered output are useful.

4. The project results are generally in line with those of other projects, like SMART.

Overall we may conclude

A. that automatic indexing involving statistically-derived information, and automatic searching, appear competitive with manual; except

B. that careful manual selection of initial search terms appears to benefit performance.