V INDIRECT INDEXING

We are familiar with a variety of keys giving access to documents. These include 'non-standard' verbal keys, like author and institution names, or journal titles, and non-verbal keys of all sorts, like size (see Cooper 1970). These may all be said to provide indirect indexing, if direct indexing refers to the use of ordinary linguistic keys. We have another form of indirect indexing when documents are characterised by their membership in document clusters. In this case documents are grouped on the basis of their initial descriptions, and the resulting classification is used to provide a new set of verbal retrieval keys. One form of non-verbal key, namely the citation, has attracted a good deal of attention as a base for indexing; and document clustering is an active research area. I shall therefore consider these two forms of indirect indexing here, to see how they compare with standard verbal indexing.

V . 1 Citations

The use of citations for straightforward literature searching is particularly associated with Garfield's Science Citation Index. More sophisticated techniques for exploiting them were proposed by Kessler 1965, for example. Research on the value of citations (outside the area of sociology of science) has on the one hand been concerned with citation matching in retrieval, and on the other with the use of citations to cluster documents. In the Project TIP (Mathews 1967) on-line search system, for example, requests may include citation specifications referring either to particular citations or to shared citation lists. There is no reason to doubt that citations are generally helpful, but not much has been done to determine their value more precisely. Salton 1971b tried to establish it experimentally by comparing retrieval with index terms only, citations only, and terms and citations combined, for the 42x200 Cranfield collection. Quite detailed comparisons involving both broad and narrow relevance sets were carried out, but the results are not very clearly presented. With the broader relevance set citations only did not perform noticeably worse overall than a manual thesaurus, and the two combined performed noticeably better. Unfortunately these tests were quite limited.

Several studies of document classification based on citations have been made. For example Schiminovich 1971 clustered about 30000 Project TIP physics documents. The results were evaluated by comparison with the document groupings induced by journal subject headings in the literature: agreement was good. Gerson 1972 reports experiments in clustering 500 patents via citations. His retrieval procedures, tested with 6 searches, were oriented to the need for full recall with patents, and he concludes that citation based clustering could reduce the effort involved in patent searching. The University of Bath 1973 has attempted to cluster journals via citations, evaluating the results by inspection.

It is not easy to get an idea of the value of citations for automatic indexing from this work. The conclusion as far as ordinary retrieval as opposed to, say, library management is concerned should probably be that citation information is worth exploiting if it happens to be available.

V . 2 Document clustering

Work in this area is briefly reviewed by van Rijsbergen 1972 and Prywes 1972.

Subject classifications are well-known devices for grouping documents. More generally, index terms classify documents: any documents with the same term descriptions form a class. If index keywords are themselves classified the resulting descriptors group documents. In the cases where keywords or derived descriptors characterise documents the number of documents in an individual class may be small, and the document grouping is not usually reflected in the physical organisation of the document file.

Bringing like documents together for searching is clearly practically convenient, particularly when document files are large. If an individual search can be confined to part of the overall collection file, substantial savings in retrieval time and effort may be achieved. In general with a clustered file, inspection of descriptions of the clusters of documents is substituted for inspection of the descriptions of the member documents. However two tier matching, first against cluster 'profiles' and then against individual document descriptions for selected clusters, may be practised. Clearly the desired economies are only attained if groups of documents contain a non-trivial number of different documents, so that the number of profiles to be inspected in matching is significantly reduced. On the other hand there should not be so few that very many documents are necessarily retrieved. Exclusive document clusters are also required, or at least duplication of individual documents should be restricted as far as possible.

The essential problem of document clustering is to ensure that the document groups coincide as far as possible with relevance classes for requests. The practical constraints on size to some extent reflect logical ones: if clusters are too small they are likely to exclude documents from the topic areas the clusters are intended to represent. If they are too large, irrelevant documents will be retrieved.

If documents are initially characterised by keywords, document clustering represents complementary processing of the same data as keyword classification. However, as noted, document clustering may also be based on other information, for example citations which have been used by Schiminovich 1971 and Gerson 1972.

Document clustering presents special problems of performance evaluation. In principle the results of searching document clusters should be compared with serial searches of the whole file, since the object of clustering is to achieve the results which would be obtained for a complete file search,

with less effort. However it is clearly misleading to say that cluster method A performs less well than serial method B when A was in fact not allowed inspection of the whole file. In general cluster methods are liable to hit a fairly low recall ceiling. So is it fair to say that method A is less good than method B because it does not exceed a recall of 60% when B achieves 90%, when A inspects only a quarter of the document file? The same problem arises when different cluster methods are compared: is a recall of 30% for a quarter of the file better or worse than a recall of 40% for a third of the file? This problem is generally referred to as the cutoff problem. It has been studied by Dattola 1969 and by van Rijsbergen 1971, 1972.

In most work on document clustering the motive has been economic, as indicated: the document classification is adopted to avoid full file searching. The hope is that though individual clusters may not contain all the documents relevant to a request, they will contain sufficient to satisfy users. However document clustering may be more strongly motivated. Clustering may be seen as a device whose primary function is to bring relevant documents together, in the hope that the set of documents relevant to a query will be more clearly separated from the remainder of the collection than its individual members are. Document clustering can therefore be described as positively as well as negatively motivated. But the success with which a classification can function as a positive retrieval device depends critically on the character of requests and relevance requirements. If document classes are not in fact relevance classes, only limited returns can be expected (see van Rijsbergen 1971, 1972, 1973).

Document clustering shades imperceptibly into file organisation in general. With large collections some rationalisation to reduce search effort is necessary. I shall not be concerned here with, for example, compressed coding schemes like that used by Thiel 1972.

V.2.1 Clustering experiments

There have been two main approaches to document clustering. One is designed to produce a one level partitioning of the document set; the other is directed to multi level or hierarchical classification. The techniques adopted for the former can be generally described as centroid methods: an initial set of cluster cores is provided, say by taking random documents, and the other documents in the collection are assigned to the classes represented by the cores according to the similarity between their descriptions and those of the core documents. Once classes of documents have been obtained they can be characterised by centroid description vectors derived from the descriptions of the member documents, for instance by selecting all the keywords occurring in at least half of these descriptions. Several iterations are usually required to stabilise the clustering, each round using the derived centroid vectors as new cores. The procedure may be refined in various ways, for example in the initial choice of cores. in some cases the specific number of classes to be obtained is regarded as important - it may be determined by external practical considerations, so the initial core set and assignment to it are strictly controlled. Other approaches are more flexible. In multi level classifications clusters of documents at successively

higher levels are determined by decreasing degrees of similarity. In some approaches individual documents with the highest degree of similarity to one another are first grouped, and further documents are added, or groups combined, at successively lower similarity levels. Alternatively, the clusters formed at one level, e.g. by centroid methods, may be treated as independent units to be partitioned. *

For retrieval, each cluster either in a single or multi level classification must be represented by a keyword description or profile. In hierarchical classification these can be obtained for each grouping by the kind of technique used to obtain centroid vectors. Searching with a single level classification simply consists of matching requests against centroids and selecting the best matching single cluster or set of clusters. With a hierarchical classification searching starts with the top node of the cluster tree, and proceeds downwards following the path indicated by the best centroid matches at each cluster level, until some suitable stopping point is reached. The set of documents below is then retrieved irrespective of further classification.

Early experiments in document clustering were carried out by Doyle 1966 and by Rocchio 1966 under the Smart project. Both used one level clusterings. Doyle did not do any retrieval experiments, and Rocchio's evaluation was relatively limited. Work in this general area, though differing in detail, was also carried out by Williams and by Ivie. J.Williams 1968a examined the use of discriminant analysis to identify terms characterising document classes in an existing classification; these key items could then be used to assign new documents to the existing classes. Experiments with several data bases, some containing more than 5000 documents, are described, but there is little information about evaluation, other than reports of good agreement between automatic and manual categorisation. A similar experiment was recently carried out by Hoyle 1973 with 124 abstracts: again there was good agreement with manual indexing. Ivie 1966 studied the use of document clustering techniques to group documents round a request during searching, in fact using citations and classification techniques of the kind described for keywords, rather than centroid methods. Evaluation was again by comparison with manual grouping.

Rocchio's work was followed by substantial investigation by the Smart project of one level clusterings. An overview is provided by Section 4 in Salton 1971a. Salton 1968b described early experiments with Rocchio's algorithm for the 35x82 and 42x200 ADI and Cranfield collections. This involved a two stage search, first against cluster profiles and then against selected document descriptions, leading to a ranked output. The results show a substantial lowering of the recall ceiling, and a noticeable loss of performance at high precision. This line of work has been carried further by Dattola. Dattola 1969 developed Doyle's technique for generating one level clusters, paying special attention to speed of classification, since this is clearly important for large collections for which clustering is really intended. The cluster method involves a great many parameters. Experiments were carried out with the ADI and Cranfield collections. Dattola attempted to measure the

---

* Terminology here is very variable: a multi level classification may be viewed either as the result of an agglomerative process leading from tree leaves (documents) to root node, in which progressively larger clusters depending on weaker similarities are formed; or as the result of a divisive process starting from the root in which clusters are partitioned into subsets reflecting stronger similarities.

amount of work involved in using clusters, and made rather misleading comparisons with full searching. The results show the best cluster sets giving a performance within striking distance of full search, at high precision; but the very wide variation with different cluster sets emphasises the difficulty of finding the right approach for particular collections, and more generally the problem of devising reliable methods of generating this type of classification. It should perhaps be noted that in Murray's 1972 experiments with Dattola's cluster method, cluster performance is noticeably less good than that of full search.

Some very small experiments with centroid document clustering for effective file organisation have been carried out by Rettenmayer 1972. Other recent work in this area is that based on citations done by Schiminovich 1971 and Gerson 1972. As noted earlier, Schiminovich made no attempt to evaluate performance in retrieval, and Gerson's tests were rather limited.

Hierarchical classification has been studied by Litofsky and van Rijsbergen. Litofsky 1969 worked with a rather crude technique not involving similarity computations but simply assigning documents to clusters by maximum keyword overlap; the clusters at any one level are independently partitioned to obtain sets of clusters at the next lower level. The procedure is strongly influenced by machine storage considerations. Litofsky's experiments were on a creditably large scale, 165x46942. The object was to generate a classification for browsing on-line, but the results were not in fact evaluated in this way. Comparisons were made with a manual classification for such properties as the number of keywords per cell (smaller numbers implying closer document relationships), and the number of nodes and documents inspected in searching. The automatic and manual classifications were very similar. Visual inspection of the automatically obtained clusters suggested they were quite plausible.

Van Rijsbergen 1971, 1972 has adopted a more formally rigorous approach to the classification procedure used, and has therefore worked with the single link cluster method (applied agglomeratively). Initial experiments were with the 42x200 Cranfield collection, but the procedure was later applied to the 63x797 Keen and 97x541 Inspec collections. A feature of the work is the attention paid to appropriate ways of evaluating retrieval performance for document clustering. Initial experiments compared cluster based searching with an idealised linear search (i.e. one with optimal cutoff). These showed cluster based retrieval could equal the ideal linear search for the Cranfield collection; but it was unfortunately inferior for the Keen and Inspec collections. In subsequent experiments with a variety of cluster based strategies performance was compared with that obtained for actual linear searches with appropriate cutoffs. The result showed that cluster based strategies could give a better performance than linear search for the Cranfield collection, and a very similar performance for the Inspec and Keen collections. These results lend some support to the view that document classification should be undertaken not merely for economy reasons, but to concentrate relevant documents.

A multi level version of Dattola's centroid clustering technique is exploited by both Murray 1972 and Kerchner 1971. In particular they explore the effects of adding new documents to an existing clustered collection.

Tests with a three level hierarchy for the 225x1400 Cranfield collection suggest that with quite simple cluster profile maintenance methods, increases of up to 50% in collection size may be allowed before performance deteriorates markedly. Kerchner also describes tests to see how cluster performance is affected by permanent changes in document descriptions following search relevance evaluation. The results show that cluster performance is not affected by document or profile modification; on the other hand the large performance improvements obtained for modification by Brauen 1969 are not maintained with clustering.

It will be evident that the performance of document classifications must be influenced by the design of cluster profiles, particularly where no serial matching against individual document descriptions for retrieved clusters is carried out. From the search point of view the cluster profile constitutes the effective index description of the documents in the cluster. The merit of a cluster is ultimately determined by the original descriptions of the documents it contains, and these are the source of its profile; but it is the profile itself which is operative in retrieval. Both van Rijsbergen 1972 and Murray 1972 investigate a range of profile definitions, the former for the 42x200 Cranfield, 63x797 Keen and 97x541 Inspec collections and the latter for the 225x1400 Cranfield collection. Both conclude that a simple approach with terms weighted by their cluster frequency is generally satisfactory.

V.2.2 Conclusion on document classification

It is unfortunate that most of the serious experiments under this head have suffered from one or the other of two defects. They have either clustered a realistically large number of documents, but not evaluated the results properly, or have evaluated the results for retrieval performance of classifying rather few documents. The real value of document clustering cannot be determined from tests with only 82 documents, for example. There is little doubt that clustering impacts recall, though this may not matter in practical contexts; however it is not clear how cluster based retrieval performs otherwise, though van Rijsbergen's results are promising. One reason for this may be the use of a theoretically well-founded classification technique, which is not typical of work in this field: some of the procedures used are so dubious they cannot really be expected to work. It is most important that the approach should be tried on a large scale.