VI MECHANISED SYSTEMS

Mechanisation at the clerical level in operational systems is not considered here.

V1.1 Standard systems

Many establishments running mechanised SDI or retrospective retrieval systems (the two need not be distinguished here) make use of standard tapes supplied from elsewhere, for example by ISI, CAS, Inspec or Medlars. These tapes typically come with a range of information including, apart from bibliographic details, titles, abstracts, and thesaurus terms, or subject headings, or keywords. Different establishments may use different selections from this information to run their local service. Particular services may also take input from more than one source: see for example Tell 1970, Williams 1972 and Hisinger 1971. I shall use the term "system" to refer to a particular SDI or retrospective search service offered by a particular establishment.

Operational mechanised systems may be characterised a) by their degree of mechanisation, and b) by the sophistication of their indexing. Many systems allow searching on a whole range of fields, see for instance Williams and Hisinger; in the present context only document or surrogate texts, and index terms, are of interest. The number of operational systems is now large, and some run on a massive scale. I shall not attempt here more than an indicative survey with illustrative references.

Limited mechanisation is represented by the use of machine-held files for searching, where the documents and requests are analysed and indexed manually. The indexing may be relatively unsophisticated, as in the Inspec use of free terms extracted from document texts (Barlow 1972); the CAC service operated by UKCIS (Barker 1972a,b) also uses keywords, and rather more extensive phrases are supplied in Project Intrex (Reintjes 1969). Alternatively indexing using thesaurus terms or subject headings from a controlled vocabulary may be supplied, as in Medlars (Austin 1968, Barber 1973), the ASSASSIN system (Clough 1971) or the GIPSY service for geological records (Moody 1972).

Full mechanisation is currently almost wholly represented in operational systems only by relatively unsophisticated approaches allowing scanning of title or abstract texts for single words or word sequences (with or without truncation). Titles are popular, for obvious economic reasons. ISI data tapes are used, for instance, by Unilever (Rowlands 1970), and title search is offered by services exploiting CT, like UKCIS (Barker 1972a,b). Abstracts are included as search fields in, for example, Williams 1972 and Hisinger's 1971 services. Full document texts are not usually available in retrieval systems, but they appear in legal retrieval services which are designed to satisfy special requirements:

for instance see Niblett 1970 and Negus 1971, and the recent survey by Myers 1973.

Apparently the only operational fully automatic system with sophisticated indexing, described earlier, is Hillman's LEADERMART (Hillman 1968, 1969, 1973). This is an on-line system with mixed data bases, and there is no real evidence on the contribution of the indexing to performance.

In general, operating mechanised systems have not been well evaluated for their retrieval performance: Lancaster's Medlars investigations, 1968a, 1969 were unusually thorough. But it must be admitted that the evaluation of large multi-purpose systems is a major problem; and it must also be allowed that many retrieval services are subject to a wide variety of external constraints which limit the choice of indexing policy and inhibit changes in it. Some of the relevant tests which have been carried out are discussed in Section VIII. It may perhaps be noted here that the Inspec adoption of free keywords after evaluation tests (Aitchison 1970, Barlow 1972) seems to be a rare instance of system response to evaluation as far as indexing is concerned.

On the whole, two approaches to running large operational retrieval systems can be distinguished. One is to endorse the use of a determinedly controlled index language, even if this is with the fatal enthusiasm of the devotee before the Juggernaut. The other is to allow maximum flexibility for searching on any or all of the range of keys naturally supplied by documents themselves, though this can be as unrewarding as picking presents from a bran tub. In any case in most systems requests and profiles are very carefully prepared to allow sufficient flexibility in the first case or control in the second.

VI . 2 Non-standard systems

Under this head I shall consider two forms of search facility for which greater user participation is assumed than in standard systems.

VI.2.1 Interactive retrieval

In the last few years interactive retrieval systems using on-line computational facilities have become fashionable: for a recent review see Bennett 1972. It is important to be clear about what this implies for automatic indexing. In general interactive retrieval amounts to no more than wrapping up old fashioned searching by the initiativeful user in shiny modern packaging. There are many requirements which have to be satisfied if the user is to be kept happy, ranging from a comfortable chair and quiet console to rapid response in presenting at least search aids like classifications or trial search output, and preferably complete search results. None of these things in themselves affect document indexing directly. But where an interactive system may indirectly affect indexing is in substituting reference to and reliance on the user for control of indexing, either in language formulation or document description.

It is perhaps useful to distinguish two types of interactive system: those in which documents may be actually inspected on-line, so iterative searching in a proper sense of the word can be undertaken; and those in which documents themselves are not available, and interaction is confined to request formulation via a variety of aids. (Of course any searching can be iterative in principle.) It may also be helpful to distinguish systems where a good deal of initiative is required from a user from those where rather little is required. Designing or conducting a search through critical use of a thesaurus comes under the first head. Some of Salton's techniques, on the other hand, require no more from a user than an indication that he does or does not like proffered documents: the request is automatically modified via the document descriptions concerned for the next cycle of searching.

It should also be emphasised that while interactive systems properly imply on-line consoles with real live users, and these are of course to be found in operational systems, useful experiments in iterative searching can be carried out by simulation: this applies to many of Salton's investigations where the live user is replaced by known relevance judgements.

The literature on on-line retrieval is by now substantial, and I shall not attempt to cover it. Cuadra 1971 contains some useful comments on experience to date. Fully interactive systems are described by Negus 1971, J.Williams 1971, Lancaster 1972b, Borman 1972, Mathews 1967, Parker 1968, Jones 1969, Reintjes 1969, Moody 1972 and Hillman 1973. These systems operate on data files of varying sizes, some by now really substantial, indexed in assorted ways. On-line search formulation is illustrated by Medlars (Barber 1973).

The evaluation of on-line systems presents considerable problems. Some systems seem to have been evaluated only by observation of user happiness: see for instance Moody 1972. This is perfectly legitimate, but may not perhaps be very discriminating. Lncaster 1972b evaluated the EARS system fairly thoroughly, using standard measures. He obtained recall and precision values of around 60% for 47 searches against about 8000 documents. In fully interactive systems performance comparisons against off-line searches are invidious, but Barber 1973 compares results for on-line search formulation with those obtained in the ordinary way. Here on-line performance does not seem to be superior to that ordinarily achieved, users conducting their own searches on-line and specialised editors obtaining comparable results.

As mentioned, the Smart experiments with iterative searching are of interest because requests are modified automatically using the information contained in retrieval samples of documents. This has specific relevance to automatic indexing. Sections V and VI in Salton 1971a give an overview of the many experiments conducted, chiefly with the 42x200 Cranfield and 35x82 ADI collections. Broadly speaking, modification strategies may be either positive or negative: that is, request terms may be promoted through retrieving relevant documents, or downgraded through retrieving non-relevant ones. The affect of this feedback information is seen by comparing performance for revised requests with that for the original ones. The results show that noticeable performance improvements can be obtained through feedback, in a fairly reliable way, with the biggest improvement typically occurring in the first iteration of several; but allowance must be made in interpreting some of the results for the misleading effects of repeatedly retrieving known relevant documents. Limited relevance feedback is actually implemented in the large ENDS system (Vernimb 1974).

The lessons of interactive retrieval for automatic indexing are not yet clear. Some systems favour natural language document descriptions, but others have adopted controlled languages. There may well be gains in user convenience with interactive systems, but not enough is known about their performance to throw light on how documents themselves should be characterised.

An interesting variation on relevance feedback involves permanent document modification rather than temporary query modification. This reindexing idea has been studied by Brauen 1969. The assumption is that queries tend to repeat one another, so that if the descriptions of documents relevant to queries are adjusted towards these queries, the documents will be preferred in future searches with similar queries. Brauen's experiments with the 155x424 Cranfield collection show material performance improvements with modified documents; however they are rather smaller for the 225x1400 clustered collection used by Kerchner 1971.

VI.2.2 Printed indexes

Printed indexes do not have to satisfy all the requirements imposed on regular indexing: they are not designed for use in machine searching. But they have to meet other requirements, like ease of use. The question here is what linguistic challenges they offer, and whether there are currently automatic systems for generating them with significant linguistic components.

Campey 1973 has surveyed index generation programs in some detail. Indexes can be divided into two (linguistic) categories, basic and ambitious. The simple KWIC index, specifically for titles, is a familiar example of the first. Considering its limitations, it is surprisingly useful. Attempts to clean it up essentially involve controlling the vocabulary. Stop lists are normal, but unrewarding words may still appear. In large indexes semantically related word forms may be far apart. Stripping and hence grouping, is the obvious way of dealing with this problem, but if stripping is done automatically it is not wholly reliable, as mentioned earlier. Given the limitations, the procedures for making these indexes and their many variants are fairly well understood.

More ambitious indexes, with word groups or phrases whose structure must be maintained, are illustrated by Armitage's articulated subject index and by PRECIS. It would clearly be nice if titles or complex subject characterisations could be automatically analysed for ordering and display dependent on the proper identification of main and qualifying words or phrases. The linguistic problems are obviously substantial. For example, titles have not merely to be broken into word groups, but some words or groups have to be selected to serve as index entries. Campey's survey suggests that currently only machine aided index production is to be found, with programs designed to organise input material which has been marked up in some way. Armitage 1967, 1968, 1970 attempted the fully automatic production of articulated subject indexes, using simple syntactic techniques, but without success, and was reduced to pre-editing of the input. It is probably the case that more progress can be made with automating the processing of manual index entries which are already strongly controlled than that of titles, which may well contain relatively uninformative components.