SECTION A : Aims, Data, Methodology

I Introduction

1 Objectives

The project was intended to carry out better laboratory tests of automatic indexing and searching techniques than have generally been carried out hitherto; i.e. these tests were intended to involve better control of variables, a more systematic choice of variables and variable values for study, comparisons over a wider range of data, and experiments on a larger scale. The tests would involve genuine material, realistic indexing and search methods, and sensible evaluation measures referring to real users.

Specifically, the object of the experiments was to establish
(a) what automatic indexing and searching techniques perform best;
(b) whether these techniques perform consistently when applied to very different requests and documents;
(c) whether they perform effectively for large document sets as well as small ones; and
(d) how they compete with manual indexing and search methods.

2 Scope

The project research has been concerned with three main problem areas. The attempt to show that automatic indexing is of value involves the determination of
(a) appropriate input data characteristics,
(b) indexing techniques, and
(c) searching procedures.
Thus under the first heading we have been concerned with the characteristics of the basic data, essentially consisting of index term lists, to which grouping and weighting procedures have been applied: thus the source of the lists, methods by which they have been obtained, their exhaustivity and the associated index term vocabulary specificity are all potential influences on retrieval performance. Under the second heading our experiments have been designed to test various statistical techniques for exploiting term occurrence information, essentially for weighting, and for exploiting cooccurrence information, in classification. Under the the third heading we have considered different matching requirements and scoring methods, with their associated treatment of search output.

Under the general heading of automatic indexing and searching, a whole range of approaches and specific procedures have been advocated in the past. We have attempted to cover the main options in a reasonably systematic way, or at least to consider representative techniques, and to try them out on sufficiently variegated data. It is extremely difficult to avoid adhocery in retrieval experiments, partly because there are too many variables for it to be feasible to investigate them fully within the scope of modest projects, and partly because available data may be limited and so preclude some possibilities. An example is provided by our largest collection, for which titles only, and not abstracts, were readily available. The set of tests conducted during the project thus has more gaps in it than the experimenters would wish.

The tests have depended on real data, i.e. on genuine documents, requests and relevance judgements, some established within a working

environment for test purposes, like the Cranfield data, and some simply extracted from an operational system, like the UKCIS material. However, the fact that our data has been obtained from elsewhere has meant that we have not been able to study human factors relating to system input directly, for example, manual indexing operations, user request formulation, etc. Equally we have studied human responses to output only via the given recorded relevance assessments, and have not concerned ourselves in detail with other aspects of user satisfaction. Some inferences about input can nevertheless be made on the basis of our tests, and observations about different features of system output likely to affect the user.

Evaluation has been in terms of effectiveness, i.e. ability of the system to produce relevant documents. The laboratory conditions of the project make nonsense of straightforward attempts to measure economic efficiency, though some remarks about probable relative operational efficiency can be made, and there is no doubt that some of the procedures studied would be economically attractive (for any system involving automation).

The type of system presupposed by the tests is geared to off-line batch searching, representing either retrospective searches or single period searches in an SDI system. No attempt has been made to investigate on-line, i.e. interactive searching, for both practical and methodological reasons. But some of the techniques investigated are obviously applicable to on-line, iterative searching.

Essentially, the project has been concerned with the tangible elements in an information retrieval system on the assumption that whatever their relative importance within a whole retrieval environment, they are clearly of some importance and it is therefore desirable that their behaviour should be properly understood.

Finally, it will be evident that as the emphasis of the whole project has been on experiments, there can have been little time for theory. But particularly since the tests were designed to be systematic, they have implied some models of indexing and searching; and models have been explicitly invoked for term weighting, as described in Robertson 1976.

3 Procedures

Given the initial test data, the experiments have all been carried out automatically; i.e. except at some input editing points, all the subsequent indexing and searching procedures have been carried out by the Computer Laboratory 370/165 computer. (Indeed, since the experimenters were not generally familiar with the technical content of much of the test data, and for efficiency words were replaced by numbers, it can truly be claimed that our investigations as a whole were objective). The general approach to an experiment, given that relevance judgements are already available for the test data, has been to apply a specific indexing or search technique to a whole set of documents and/or requests, to conduct a complete* scan of all the documents for all the requests, and to derive an average performance figure for the requests, using the relevance judgements.

* logically, not necessarily physically

II Test Data

1 Form

1.1 General Characteristics

The test material (document, request and relevance judgement sets) was intended to satisfy basic requirements for experimental validity: i.e. it should be

(a) real material, not simulated, artificial, or created by the project itself, but emanating from genuine or at least quasi-genuine indexers and users, independent of the project. It should preferably derive from an actual retrieval service, or if not, should be comparable with service material. This requirement was not difficult to satisfy in principle since the project workers were not in a position to generate material, and some existed outside. However, it is not easy in practice to obtain material which is wholly satisfactory, since other projects may have quite specific objects in setting up test data, which limit its utility for subsequent projects; and we indeed found ourselves restricted in this way.

(b) adequate material, consisting not of a single small set of documents etc., but of several sets allowing comparisons between indexing and search techniques, and of sets large enough to ensure statistically reliable results. The requirement for variety was not difficult to satisfy, but that for large scale sets was, and only one such set, of a limited character, could be obtained and processed in the time available.

As the different data sets came from outside the project, it will be evident that while the test material satisfied the general need for variety, the particular values for environmental variables or factors represented a pretty arbitrary subset of those deserving study: for example subject, vocabulary hardness, expertise of indexers, users etc. The project itself was of course not able to introduce any control here, so it is possible, though we believe it unlikely, that our test results are really due to unexamined, rather than the examined, features of the material. Some input variables particularly relating to indexing and searching were as systematically studied as was feasible, as described below. In general we feel that though the hand of Cleverdon is directly or indirectly visible in most of the material used, in that the Cranfield project influenced the way it was set up, the actual data differs so much in substantive properties like subject, or formal ones like numbers of documents, that results which are consistent over different data sets are reasonably reliable.

A subsidiary requirement was that the material shoud be

(c) informative material, through having been used by other research projects. Its further use would then extend the range of comparative experiments carried out by workers in the field in general, and thus justify more comprehensive conclusions than individual projects are typically able to draw.

1.2 Material and Collections

The phrase "test collection" is frequently used to refer to a set of documents, a set of requests, and at least one set of relevance judgements for the requests. A "test collection" thus consists of the essential raw data, the documents themselves, the expressions of user need or interest,

perhaps presented as written statements, and the judgements. The latter are sometimes based on the documents themselves, sometimes on rperesentatives like titles or abstracts; and they sometimes refer to the need underlying the written statement, sometimes only to the statement. For the purposes of the Report, to avoid ambiguity, the phrase "test collection" with the meaning just discussed will be replaced by the phrase raw material.* This will be deemed to refer to written items only, i.e. documents and requirement statements, and implicitlyto judgements relating to the latter rather than to undetermined underlying needs.

A further distinction is appropriate where indexing or judgement is based not on the raw material itself but on some representative or reformulation of it, like abstracts for documents, or amplified need statements. That form of the raw material which is actually input to indexing and judging will be called the source material (or source for short). This may or may not be identical with the raw material; indeed the distance between raw and source material may differ for document and request indexing, while the form of the request to which judgement refers and that of the document to which it applies may differ from those used for indexing. In general in our experiments, raw and source material have differed for document indexing, and have been the same for request indexing; while the judgements have sometimes not been applied to the same source form of the documents as has been used for indexing, though they usually refer to the request indexing source.

Source material as such has not been used directly for our main experiments. These have been based on descriptions constituting primary indexing characterisations of documents and requests (accompanied by simply coded judgements), which may or may not have been further modified by a variety of automatic indexing procedures, and which may be exploited by a variety of search strategies. The primary indexing itself consists essentially of simple word lists. A version of a given body of source material represented by primary index descriptions of some kind will be called a test collection, or collection for short. Different sets of descriptions derived from common source material using different initial indexing methods, for example manual and automatic extraction from the document texts, constitute different versions of the source material. In turn, as different source material may stem from the same raw material, we may in principle have several sets of versions. Comparisons between collections which are different versions of the same source or raw material are important for some purposes. But equally comparisons are required between collections derived in the same way from different raw material. In particular, in the context of the project, whether the primary indexing is in what may be called manual or automatic mode is of some importance.

Every attempt has been made to create different collections for single bodies of data, i.e. different versions and version sets for this raw material, in a sufficiently consistent way, and equally to generate collections for different bodies of raw material according to the same principles. But inadequate resources have necessarily meant that we have not been able to do this in a really satisfactory way: we have not been in a position, for example, to supply initial manual indexing of technical material, or to keypunch large amounts of material for automatic processing. Thus the range of collections representing different versions of the same source is typically very limited, and this also applies to collections derived from one body of raw material: our primary indexing has tended to differ in both mode and source (though for each mode there has been virtually no

* Underlining introduces a defined use of an expression, which will hopefully be consistently maintained in the rest of the Report.

variety in specific method of primary indexing). The term version will therefore be used more loosely than above, to refer to collections derived from the same raw, even if not the same source, material. We nevertheless hope that there are sufficient connections between the different test collections for proper comparisons of retrieval results.

Some subsets of documents (perhaps with subsets of requests), defined by substantive rather than formal properties of the raw material, have been selected by the original compilers of the data. A well-known example is the Cranfield 200 set of aerodynamic documents selected from the larger set of 1400 documents on aeronautical topics. Since these subsets have some distinctive properties, they have been regarded as distinct bodies of raw material, generating collections distinct from those derived from the larger host bodies. Some comparisons between results obtained for full and subsets are thus in order, but others require totally different raw material. Conversely extrapolation from the smaller to the larger inclusive sets requires care. Particular experiments may have involved special selections of requests and documents, on formal grounds, and in particular random selections from, or divisions of, a document set were required for one series of tests. Such selected sets of items are treated as subsidiary collections and are detailed at the appropriate point.

A few other subsidiary collections, defined by slightly modified or different primary indexing, were also set up for specific reasons. Details are indicated below.

The details of the raw material and collections are given in paragraph 1.4 below; they are summarised in Figure AII.1. For immediate convenience this important information is repeated here: the table below lists the collections, gives their size in terms of numbers of requests and documents, and indicates the essential character of their primary indexing. It will be seen that the collection name constitutes a schematic summary of this information: for example, C200A represents the Cranfield set of 200 documents automatically indexed from abstracts. It will also be evident how far the collections derived from a given document set differ in both mode and source of the primary indexing.

raw/source material name	collection name	size reqs	size docs	primary indexing
Cranfield	C1400I	225	1400	manual from documents
	C1400A	"	1396	automatic from abstracts
	C1400T	"	1400	automatic from titles
	C200I	42	200	manual from documents
	C200A	"	"	automatic from abstracts
	C200T	"	"	automatic from titles
Inspec	I500I	97	541	manual from abstracts
Keen	K800I	63	797	manual from abstracts
	K800T	"	800	automatic from titles
	K400I	47	404	manual from abstracts
	K400A	"	407	automatic from abstracts
	K400T	"	408	automatic from titles
UKCIS	U27000T	182	27361	automatic from titles
	U27000P	"	"	automatic from titles via profiles

UKCIS relevance judgements

The UKCIS material differs from the rest in having incomplete relevance judgements. The original relevance assessments were made for the pooled output of the UKCIS project searches on the different fields (titles, keywords, digests) available for searching in the service, possibly with slightly different forms of the profile. The average number of documents thus assessed per profile is 234.0, representing a small proportion of the total of 27361, with an average of 58.9 relevant. The UKCIS project searches all involved Boolean search formulations of some complexity. The problem therefore arose, for our project, of evaluating performance for any search procedures of a different kind, retrieving different documents: for any documents retrieved and not already assessed would automatically be deemed non-relevant, even though some of them might have been judged relevant if offered for assessment.

Our procedures for dealing with this problem are described in Paragraph A III.3.

1.2.1 Alternative requests

Some experiments were concerned with the effects on performance of variations in the primary request indexing derived from the source need statement, in relation to one form of document indexing. Examples are comparisons between request indexing in manual and automatic modes, to provide specifications for searching automatically indexed documents; between more and less careful indexing in a specific mode, say between simple and carefully thought out manual request term lists; and between structurally different indexing in some mode, say between coordinated terms and more complex Boolean formulae. A given collection version of the documents may therefore have more than one applicable set of primarily indexed alternative requests. A specific request set may of course also be applicable to more than one version of the documents: for instance, given indexing by simple extracted words, manually indexed requests may be applied to manually or automatically indexed documents. (In such cases there may be trivial differences in the request specifications since the document indexing vocabularies may differ). The treatment of requests and documents may in principle be strongly correlated: an example would be in the use of a controlled precoordinate term vocabulary. In the test material used by the project there were only weak correlations, the indexing of the documents sometimes imposing constraints on that of the requests in that request terms representing word stems were determined by the particular character of the document word groups defining stems, or that of requests contstraining documents as, for example, when arbitrary request word truncation implied full natural language word document indexing.

Where alternative requests exist for a given version of the documents, one of the alternatives, usually that first generated and by procedures analogous to that used for the documents, has been regarded as the main one.* The different combinations of alternative requests with one version of the documents are not taken as defining different collections. Alternative requests for a given document set are simply listed along side the main ones. In the various figures the main and alternative are labelled by their indexing mode which may be 'm' for manual or 'a' for automatic.

The whole question of such request document relationships is a complicated one, and increasing emphasis was placed during the project on the treatment of requests rather than documents. The points involved are

* The Keen 800 titles are anomalous in having manual requests only.

more fully discussed in the description of the experiments. Here it is sufficient to note that the alternatives investigated represent only some of the possibilities.

For some collections, a more careful manual indexing of the requests was available as well as a rather crude one, representing either a more cautious selection of terms from the given need statements, or the addition of plausible terms. These alternative requests are labelled 'g' for good in figures.

1.2.2 Relevance variants

Some of the research projects from which material was obtained allowed different grades of relevance in the relevance judging (non-relevance is not regarded as a grade). Thus the Cranfield project allowed four grades, Inspec and Keen two. Relevance need is an important environmental variable, and since the necessary information was available, some tests were repeated with different sets of relevance judgements representing different needs. The Inspec and Keen material distinguished strongly and weakly, or highly and partially relevant documents, and some of the tests with collections derived from this material were carriedout with the subsets of strongly relevant documents only, as opposed to the complete sets of all documents deemed relevant in some degree. For the Cranfield material relevant documents in the highest two grades were grouped to form highly relevant sets, as opposed to the full sets covering all four grades. As it is the case that a non-negligible number of requests have no highly relevant documents, tests involving the latter only forced the selection of a subset of the requests. However, since the tests with the restricted relevance sets were essentially checks on the main experiments, rather than of major concern, the use of a request subset was not deemed to imply a distinct collection and details of the relevance sets are therefore simply listed where appropriate for the regular collections. Further, the fact that only by including partially relevant documents could all the initial requests be covered meant that the full sets were always taken as the primary ones.

For some of the Cranfield material three quite distinct full relevance sets were established by distinct judges, originally for the purpose of validating the Cranfield project findings. One of these sets seems to have been based on much the same narrow interpretation of relevance limits as the initial full set, while the others were broader, giving larger sets for each request. As these sets were used for some experiments, the data descriptions includes details of them as alternatives to the main sets.

For reference the various sub and alternative sets of relevance judgement will all be called relevance variants, the context making it clear which type is intended. Variants representing highly relevant documents only are labelled 'H' in figures; the Cranfield alternative sets are named "A", "B" and "C"; most of our tests have been with one of the broader sets, B.

1.3 Test Material Processing

The raw material supplying our test collections had been indexed in one or more ways by the original projects, and one form of the initial project indexing was taken over to provide a link between earlier and later experiments with the same material. The form selected usually represented simple manual mode indexing of both requests and documents generating extracted keyword lists; but for the UKCIS material the original indexing provided Boolean profiles using word fragments and strings intended for

searching the actual title texts. Except for this UKCIS data (leading to collection U27000P), for which it was inappropriate, the manual keywords were processed to identify common stems which were adopted as terms and replaced by numbers for easy manipulation. Details of the comprehensive processing and standardisation to provide the computer data files in a regular format actually used are given in Section D, where our data management techniques are described. This includes a note on the stemming procedure: for the present it may be noted that this involves picking up word groups in the vocabulary derived from the document set, to which request words are assimilated.

Where possible the test material was also indexed in the automatic mode to obtain keyword and hence term lists for titles and for (titles+) abstracts. These were set up in a systematic way by deleting ordinary 'stop' words like prepositions and conjunctions from the given texts and then processing the remaining word lists in the same way as the manual ones. For the abstracts the frequency of occurrence of each term in an abstract was incidentally obtained.*

These term lists constitute the primary indexing for the collection which their generation serves to define. Thus for the Cranfield material, as Figure AII.1 shows, the complete set of 1400 documents provided three collections representing automatic indexing from two different sources and manual indexing from a third. We do not feel that the replacement of actual word forms by stems requires any justification: insofar as its effect on retrieval has been tested, e.g. by Salton (1968a) and Cleverdon (1966), it seems beneficial, and is certainly practically convenient in reducing the size of the term vocabulary used.

Occasional errors in the data processing meant that there are some small differences between collections derived from the same material: thus the odd document has been lost or gained, accounting for the discrepancies in numbers of Figure AII.1. These have been disregarded as without effect on the experimental results. Other negligible oddities are due to the fact that document subsets were sometimes separately processed, and not simply selected from larger sets.

There are also odd numerical differences between the data as we have used it and as it was originally used: an example is the UKCIS material where we have 182 profiles instead of the original 193, since we rejected some of the latter as having either no assessments or no relevant documents. Such small differences have been regarded as irrelevant in comparisons between our results and those obtained by previous projects using the same material.

Two subsidiary collections are associated with the primary processing for some of the data. For the Cranfield 200 abstracts, a word form version C200Aw, representing the texts without stop words but before stemming, was retained to check the value of stemming. The Cranfield 200 manually indexed collection, C200I, was set up some ten years ago using the elaborate descriptions involving partitioning to form "themes" and interfixing for "concepts" constructed at Cranfield. At the time a decision was taken to remove terms occurring in only one theme. As this form of the data was used in subsequent research, it was retained for the present project to maintain compatibility. However, to check the effect of the term elmination, the deleted terms were reinstated for the subsidiary complete collection C200Ic.

* The Cranfield abstract collections were in fact obtained already processed in a similar way from the SMART Project.

U27000P

The UKCIS material generating collection U27000P was very different from the rest and could not be processed in the same way. It indeed presented a variety of challenges to our whole view of automatic indexing and searching.

Specifically, the relevant data was supplied in the form of document title texts and search profiles set up by UKCIS in their standard way. The profiles differed from any requests we had previously used, or created in our regular way, in having a Boolean structure, i.e. a structure involving more than mere coordination; and, more importantly in the present context, in involving user-defined word stems and fragments, and also multi-word terms. In our ordinary practice stemming is fixed by the document indexing, and request words are treated to match. Multi-word terms have not ordinarily been treated in automatic classification and weighting experiments, and indeed classification may be regarded as a method of identifying them; but in principle such statistical techniques can be applied to any multi-word terms established at the time of document indexing and treated as units thereafter. With the UKCIS data stems, or rather fragments, since any part of a word may be used as a term, and precoordination, are determined at search time by the user. Different users may thus treat a word, set of words, and string of words in different ways, which is not possible with prior stemming and precoordination. The UKCIS system is extremely hospitable since front and/or back end truncation of any word or word sequence is allowed, which is of course natural for chemical compound names.

It must be recognised that such request-based approaches to indexing present substantial problems for any procedures, like term or document classing, which exploit information about the co-distribution of index terms in documents. Weighting techniques relying on the distribution of terms would be easier to manage, but computing weights to organise output only after document searching provided the necessary distributional information would clearly be awkward. For test purposes, however, where the request set is fixed, all the necessary distributional information can be obtained for the request terms.

The consequence of this property of the data is that complete collection information of the kind usually obtained, and given for the other test collections below, cannot be provided for the U27000P collection. Thus the number of different words in the titles can be given, but not(in any useful way) the number of potential index terms. Most of the details for this collection refer specifically to the set of index terms defined by the requests. More particularly, terms are strictly associated with individual requests, and have identifying numbers local to requests; the fact that the same term may occur in different requests is thus disregarded.

It is further the case that the UKCIS profiles are of two rather different kinds: those which have a strict Boolean logic, and those which have a complicated mixture of weighting and Boolean logic. In the latter the terms in a parameter or group of terms linked by 'or' are ordered, and the weight of the first matching one is taken. If the sum of weights for those groups which match exceeds a threshold, a document is retrieved. As with the ordinary Boolean requests a document may be rejected if it matches a term in a NOT parameter ; but there is no requirement that every group shall match, i.e. that the overall 'and' structure of the profile be satisfied. The complexity arises because terms may have negative

weights and are not ordered within a parameter by weight e.g. from highest positive to lowest negative.* We felt that such profiles could not be used in any way other than that intended, and so our experiments with structured profiles have regrettably been mainly with the 75 profiles originally having a strict Boolean structure. The resulting subsidiary collection, representing a request but not document subset, is the U27000Pb collection.

1.4 Data Details

Figures AII.2 - 8 give the essential facts about the test data, i.e. raw material and derived collections, used in our experiments. Information about the raw material is summarised in Figure AII.2. As all of it was obtained from elsewhere, full information about the way the documents were selected, manually indexed if this was done, about the way requests were supplied and indexed, and about how relevance judging was carried out, should be sought in the source project publications listed. In our project we found that most of the collections had some defect for the purist in experimental methodology: but we had no real alternative but to take what we could get. The problem of incomplete relevance judgements was a particular difficulty with the UKCIS material, where these were obtained only for the small output of the original project searches. Notes on such germane points are given in the final comments section for each body of material. Figure AII.3 summarises the word processing involved in deriving the document and request primary indexing of the collections from the raw material.

Figure AII.4 provides information about our actual collections as characterised by the primary indexing. This is analysed further in paragraph 2.1 below. As noted above, the U27000P collection does not have primary indexing in the ordinary sense, so details for it are given separately in Figure AII.5. This figure also contains details about the original assessed output.

A comparative table, Figure AII.6, summarises the main numerical data for the collections used in our tests, in terms of the primary indexing. This clearly brings out the very varied nature of our experimental material. Details of subsidiary collections, where they differ from the main ones, are given in Figure AII.7. The following Figure AII.8 gives some information about alternative requests and relevance variants.

2 Properties

2.1 Totals and Averages

Since the formal properties of the test data must have some effect on performance (indeed one object of the project was to identify critical formal properties), numerical values for a variety of collection properties are listed for all the test collections in Figures AII.4 and 5. These show the distribution of relevant characteristics for the different retrieval entities, documents, requests, relevant documents and terms, following a common format.+ Thus for documents indexed by terms we have maximum and minimum numbers of terms per document, the total number of terms represented, the total postings, and average postings per document; for the relevant documents (i.e. inverted relevance judgements) we have the maximum and minimum number of requests for a relevant document, the total number of different documents represented in the relevance judgements, their total

* This type of profile is apparently no longer used by UKCIS.

+ In the language of our standard data formats, described in Section D, the

'postings' and the average number of requests per relevant document; and so on. These details are laid out in the tables so that the corresponding information for different collections derived from the same raw material can be compared (horizontally) and that for collections derived in a similar manner from different raw material can also be compared (vertically). Thus we see that for the Cranfield material the number of index terms per document for the C1400I, C1400A and C1400T COllections respectively is 29.9, 53.6 and 7.8, while compared with the C1400T collection the number of terms for the K800T AND U27000T collections respectively is 5.5. and 6.6.

The tables show how much variety there is within a collection: for example, the shortest document description in the K400A collection has 11 terms, the longest 170; in the U27000T (and P) collection, one request has only 1 relevant document, while another has as many as 554. When comparisons are made between collections, it is evident that there are many formal differences. Apart from the obvious ones of size in terms of numbers of documents and requests, which range from 200 - 27361 and 42 - 225, the size of vocabulary ranges from 459 terms for C200T to 17537 for U27000T; while the term numbers for C1400I, C1400A and C1400T are 2683, 4949 and 1175 respectively, compared with 939 and 987 for K800I and K800T, or 17537 for U27000T. Averages also vary, for instance for terms per document from about 5 to over 50, and for terms per request from under 5 to over 12. The number of relevant documents per request is naturally larger for the UKCIS collections, but is larger for the Keen 800 document sets than for the Cranfield 1400 ones. It is interesting that the proportion of the document set relevant to some request is as high as nearly 28.8% for the UKCIS collections, though it is much higher for the smaller collections. The proportion of the U27000T term vocabulary figuring in the requests, on the other hand, is very small. The different ratios of term numbers to document numbers also deserve comment: 651/408 and 987/800 for K400T and K800T are fairly predictable; but 17537 terms for 27361 U27000T documents presumably reflects the fact that the UKCIS titles contain many specific chemical names. Finally, the number of documents per request term is of considerable interest: the table shows clearly how the request terms tend to be more frequent ones. Thus the average request term will by itself retrieve 179 documents from the C1400A collection and 285 from the U27000T collection.

Figure AII.8 provides analogous information for the alternative requests and relevance variants. The latter in particular show how few highly relevant documents there typically are.

2.2 Frequency Distributions

A more detailed picture of the test data is given by the frequency distributions of characteristics for entities. The distribution of terms for documents, terms for requests, and relevance judgements for requests is normal across all the collections, and hence all the raw material. That for the inverted lists is equally consistently 'Zipfian'. The full information is too voluminous for the Report, and it is therefore illustrated by the data for the C1400I and C1400T collections, for manual and automatic indexing respectively, in Figures AII.9 and 10.

3 Relationships

For experimental purposes, where compete data consisting of a fixed

file of requests, documents and relevance judgements is initially available, it is useful to consider some of the resulting relationships between these entities, as they must limit the potential scope of the indexing and searching strategies to be investigated.

3.1 Matching Relatinships

We have found it instructive to consider the simple request-document matching relationship determined by the primary indexing of the collections. Thus we may compare

(1) the average number of given, or 'starting' request terms for a collection;
(2) the average number of 'retrieving' terms per request, where the number of retrieving terms for a request is the maximum number of its starting terms matching some document;
(3) the average number of matching terms per document retrieved, over the set of requests; and
(4) the average number of matching terms per relevant document retrieved.

The relevant details for the collections are given in Figure AII.11. These all show that the best matching scores per request are generally much lower than the possible scores. They also (comfortingly) show that the matching scores for relevant documents retrieved and for all (generally non-relevant) documents retrieved do differ. However, since in some cases, e.g. U27000T and most of the Keen collections, the scores even for relevant documents are very low, the difference between relevant and non-relevant scores is, in real terms, very small.

Of course the object of retrieval experiments is to do better than some such baseline as that represented by these figures; and they are not wholly straightforward, as appears when the seemingly comparatively good K400A figures are compared with performance measured in terms of recall and precision, shown in Figure AII.12. The scores presented here are nevertheless useful in bringing out some essential properties of the test data and hence the challenge to be met.

3.2 Relevance Relationships

Another view of the structure of the test data is obtaining by considering the discrimination achieved by the primary indexing between relevant and non-relevant documents.

3.2.1 The Cluster Hypothesis

Retrieval assumes some distinction between documents relevant and not relevant to a request; or, to put the point another way, that the documents relevant to a request are more like one another than they are like non-relevant ones. This assumption has been named the Cluster Hypothesis. The extent to which it holds for given data can be indicated by computing the association, based on their index descriptions, between relevant documents, and between relevant and non-relevant documents, for each request. Aggregating the reuults over the set of requests gives a characterisation of the test data as a whole, showing the relative distribution, over the range of association coefficients, of relevant-relevant, and relevant-non-relevant association values. The two distributions are conveniently displayed in the form of a histogram relating cumulative percent of associations against value. The association coefficient used is normalised symmetric difference, defined for a pair of documents A and B with binary index descriptions as

$$\frac{|A \Delta B|}{|A| + |B|} \quad \text{or} \quad 1 - \frac{2|A \cap B|}{|A| + |B|}.$$

This Cluster Hypothesis Test is discussed in van Rijsbergen 1973 and 1975a.

Results of the Test for our experimental collections, using their primary indexing, are given in tabular form in Figure AII.13, and for some relevance variants in Figure AII.14. Figure AII.15 provides a graphical illustration for the C1400I and K800I collections. (Note that some of the details differ slightly from those published in 1973, which were obtained with a faulty program. The original observations on collection differences are, however, supported by the later correct results). As calculating large association matrices is extremely expensive, the Test for the U27000T collection was applied to a matrix for a random 5% of the documents. The Test could not be applied to the U27000P collection, as it cannot be used in a straightforward way with request-based indexing.

The Cluster Hypothesis Test figures for the regular relevance judgements show both how collections derived from the same raw material differ, and how those derived in a similar manner from distinct material compare. In general, the separation between the relevant-relevant and the relevant-non relevant distributions is substantial for the Cranfield collections, and much larger than that of the Keen collections, which is poor. The Inspec collection separation is moderate. The different shapes of the distributions reflect average index description length: with short descriptions there are many total dissimilarities (zero similarities), leading to high percentages in the final column of the histograms. For the Cranfield data, the title based collections perhaps show somewhat smaller separation than those based on other indexing sources, but separation for the K400A collection is poor compared with that for the Cranfield abstract collections. The test results for the relevance variants parallel those for the regular judgements, though for individual collections there is a slight tendency towards better separation for highly relevant documents only.

These observations are informal, and the graphs are indicative only in the absence of formal measures of separation. We have not been able to supply the latter, and have therefore taken the precaution of adopting a rather conservative view of the data when drawing conclusions from it. Our simple analysis is summarised in Figure AII.16.

The Cluster Hypothesis Test can naturally be applied to any form of document indexing which readily lends itself to the computation of association coefficients. It can then be used to show, not only how radically different primary indexing affects document relationships but, for example, how changes to an indexing vocabulary may alter the document relationships, or how more or less exhaustive indexing can affect them: one possibility with automatic classification, for instance, would be to check the effects of adding class related terms to initial document descriptions. The main practical difficulty about this is the computational effort of forming and examining the large association matrices involved, and we have not continued further in this direction. In any case, our main project work has been on the manipulation of requests, and requests are not directly involved in the Cluster Hypothesis Test. The value of the Test for work on automatic indexing techniques designed to improve retrieval performance through changes in request treatment or search strategy, is in providing a view of the initial data which may be compared with and explain (to some extent) actual retrieval results. More generally the Cluster Hypothesis

exhaustivity is comparatively well understood.

In a test context like ours, not merely are environmental parameters determined, but many system variables also. If data obtained from elsewhere has been manually indexed from abstracts, say, it may not be feasible to obtain comparable indexing for full document texts. Thus it may be difficult to determine at all precisely how environmental parameters or other system variables are affecting the behaviour of the variables being studied. Perhaps the most that can be said is that provided the bodies of test material are variegated enough with respect to factors not open to direct investigation, it is likely that consistent behaviour across the test collections in a variable being studied is attributable to the values assigned to the variable rather than to other system components.

Experimental research is of course intended to throw light on the effect of environmental parameters by cross-system comparisons. But it is impossible for a single project to investigate such parameters comprehensively. It is equally hardly practicable for a single project to study many different system variables in any detail. Even investigating those variables which are the focus of study in a sufficiently comprehensive way is difficult enough. We hope that the runs we have done
(a) reflect a sufficiently large range of contextual parameter and variable values, and
(b) test significantly different values of the system variables under study.

It is convenient to divide system factors relevant to indexing into three classes related to system operation. Thus we have input factors affecting the documents and requests presented for indexing, including environmental parameters like subject and user need, and system variables like indexing source and mode. Indexing factors cover the types of information supplied for documents and requests, primarily reflecting choices for system variables like indexing vocabulary, but also environmental constraints like the vernacular of system users. Output factors influence the use made of the indexing information supplied, including system variables like mechanical search procedures or request document matching functions and environmental parameters like physical output limits.

As this division corresponds to phases of processing, in practice choices taken in later phases tend to be constrained by decisions taken earlier. The difference between input and output factors in relation to indexing is that the choice of indexing technique is to a large extent independent of input choices, whereas output choices may not be necessarily or usefully independent of indexing. The separation of indexing and searching is indeed rather artificial just because the way indexing information is exploited in searching tends to follow from the form of the indexing. But the same information may nevertheless be exploited in different ways, if not in any way, so searching is treated separately.

The next three sections discuss the three classes of factor, and input and output factors in particular, in more detail. They relate the rather general notion of factor to our project data and objectives. Thus our concern with automatic indexing and searching techniques based on natural language has meant that the factors selected for study have been categorised as input, indexing or output factors in a way which seems appropriate to systems based on such techniques. In the following sections relevant environmental parameters for the test material are noted and given values listed. The factors selected for experimental study,

primarily system variables, are then indicated, along with the value sets or groups of individual variable values we have examined. That is, for the variables with which we have been concerned we refer at a higher level to a value set as a group of very similar variable values, and at a lower level to specific values. For example, if indexing exhaustivity is deemed a system variable, we may refer to a high exhaustivity value set which in turn covers specific high indexing values of, say, an average of 40 terms per document, or an average of 50. The paragraphs below summarise the project variables and their value sets. The specfic variable values investigated are fully described in the account of our experiments in Section B.

2.1.1 Input factors

Input factors include those environmental parameters and system variables characterising given documents, request statements and relevance needs; for our project those directly affecting the provision of descriptive information constituting primary indexing are particularly important.

Input factors may be distinguished as global and local; or perhaps, in terms of their likely effect on retrieval performance, as remote and immediate. Global remote factors include environmental parameters like document and request subject area, conceptual hardness, document or request type and level, language and technical terminology, temporal range, and so on, and user characteristics like competence; remote system variables include a variety of basically economic factors. Local/immediate factors include environmental parameters like user relevance requirements (strong, weak and selective, exhaustive), and system variables like indexing source and mode, and also, for our project, features of the primary natural language document and request descriptions.

In general, as noted earlier, comparisons between runs on collections derived from different raw material will allow for variation in global factors (each body of material providing an aribtrarily selected bundle of parameter and variable values). Equally, comparisons between runs on collections derived from the same raw material will allow for variation in local factors. Figure AIII.1 summarises global input factor properties of the test data along the lines of Sparck Jones 1976b: the details in fact concern environmental parameter values only, since the original research project genesis of the material inhibited systematic economic behaviour. The figure clearly shows that the bodies of raw material are very different in character; so while they do not really permit well-organised comparative studies of any positive effects of global factors on performance, negative conclusions about the absence of effects may be drawn from our tests.

Our research made some allowance for some local environmental parameters representing relevance requirements. The relevance variants available for some of test collections could be treated as embodiments of requirements for strongly or weakly relevant documents, and in the treatment of output user desires for any or all relevant documents may be hypothesised. These points are more fully treated later.

The local system variables selected for investigation as directly relevant to automatic natural-language based indexing, and the value sets studied were:

(1) indexing mode : manual or automatic
(2) indexing source : title or abstract or text
(3) indexing description exhaustivity : low or medium or high
(4) indexing vocabulary specificity : low or medium or high

2.1.2 Indexing factors

In principle there are remote environmental parameters which may affect indexing: examples might be constraints imposed by extensive national or international use of index descriptions; and there are clearly economic system variables relating to indexing. Immediate environmental parameters are perhaps illustrated by requirements for index description comprehensibility, for instance demands for transparency, brevity or systematic structure in descriptions. System variables naturally relate to all aspects of the index language itself and those of request and document descriptions resulting from its use.

Environmental parameters affecting our test data are not particularly obvious, and we have not considered costs other than incidentally. There is little point here in discussing all the variables characterising indexing of kinds totally different from that studied. Indexing languages and descriptions in all their manifestations have been extensively treated in the literature. There are indeed many alternatives available even within the relatively restricted scope of the approaches to indexing we have studied. These are best considered in the context of the detailed discussion of the procedures in Section B. Here it is sufficient to note that, broadly speaking, the system variables and value sets appropriate to natural-language indexing which were directly investigated were:

(1) term classification : a) type - tight or loose
b) use - for substitiution or addition
(2) term weighting : a) type - document or collection or relevance
b) use - ordering or selection

2.1.3 Output factors

We have chosen to assign searching to the system output component, so output factors are those relating to the search operation and its products. But as noted, it may not be easy to separate indexing and output factors with respect to the detailed operation of a retrieval system, and in particular in relation to the treatment of requests. Our categorisation of factors is thus mainly intended to act as a device for studying the comparative effects on performance of specific ways of exploiting specific types of information.

In general, for output, global environmental parameters can be identified relating to user constraints on search procedures and acceptance of the form and volume of search output. Economic system variables are again relevant, though we have not considered them explicitly in much detail. Immediate environmental parameters perhaps include user requirements for well-defined search operations and well-structured output.

The test data does not refer to such environmental parameters directly; and the general approach adopted in experiments with automatic indexing and searching techniques has been to carry out exhaustive searches with full output which is typically evaluated with rather abstract performance measures. The assumption seems to have been that such an approach is required to provide a picture of system performance in all circumstances.

It was natural with early small test collections, but is clearly of limited value with large ones. We have followed this course ourselves, but have recognised the importance of environmental parameters and have hypothesised some by, for example, applying search output cutoffs, as described later.

With respect to the immediate system variables related to automatic indexing and searching, we found ourselves dealing with complex and very different approaches to searching. This variety is partly a function of the test material itself, and partly associated with comparisons between our own work and that of other projects. We further found ourselves concentrating increasingly, in the course of the project, on request characteristics and their implications for searching, rather than on document characteristics, as the most important influence on retrieval system performance. The UKCIS SDI profiles of the U27000P collection in particular represent an extreme case of emphasis on the search component on the system, with document indexing via request indexing embodying a range of matching options. As mentioned in Chapter II, the documents of this collection are not independently indexed in any strong sense. A decision has affectively been taken to treat the document title texts as sufficiently indicative and non-restrictive document index descriptions, which are not subject to further analysis. The documents are further indexed by the requests not merely in the ordinary sense that any document retrieved is indexed by the search selecting it, but in a rather stronger sense that the apparent document descriptive entities, the title words and word strings, may be replaced in different ways, for different requests, by other entities i.e. fragments or substrings. Further, the requests, unlike those of the other collections used, explicitly incorporate complex search strategies.

The attempt to compare very different types of request and search operation in relation to automatically obtained or supplemented indexing information forced us to break searching down into a number of elements. These are the way in which the document set is inspected: not all strategies allow for inspection of every document which might have something in common with a request; the way in which individual documents are viewed in relation to a request, i.e. how document terms are selected in matching; and the way in which an actual matching score is computed for a document. These system variables, and the value sets for them we have worked with, can be listed as follows:

(1) scanning strategy : all documents or some documents
(2) matching condition : any terms or some terms
(3) scoring criterion : equal terms or unequal terms

For convenience here it may be noted that the natural output of searching procedures may be minimally, partially, or fully ordered. In the first case all the documents retrieved have the same rank; in the second the retrieved documents may have different ranks, with more than one document per rank; while in the third there are as many ranks as documents. However, output naturally generated in one form may be forced into another to facilitate comparisons: thus for some purposes it is convenient to obliterate differences of ordering while for others it may be useful to establish a full ordering; by distributing n documents of the same rank randomly over n different ranks. For simplicity we may refer to minimally ordered output as unordered, and to partially or fully ordered output just as ordered.

Ordered output as just described refers to matching documents. A more comprehensive view regards an entire collection as ordered by a search, with non-matching documents in lowest rank. This is attractive for evaluation since retrieval output is the same size for all requests. But when such a complete ordering is also fully ordered non-matching relevant documents must be assigned specific ranks. We distribute these documents randomly over the available rank positions, and Salton adopts a similar procedure. It must, however, be recognised that such pseudo-retrieved relevant documents obscure the genuine matching picture.

As the original ordering of search output may be changed, we distinguish output form and output type: the former is that naturally generated, while the latter is that version used for performance evaluation and comparison. In general form and type will be the same, but they need not be and changes are required for some experimental purposes.

For convenience, the system variables and value sets we have investigated in the project experiments are brought together in Figure AIII.2.

2.2 Baselines

For test purposes, it is desirable that some retrieval performance baseline should be established over which improvements derived from different indexing or searching techniques may be sought. No particular form of indexing and searching constitutes a necessary baseline: one way of choosing a baseline would be to select the simplest form of indexing and searching; another, appropriate to data associated with different projects, is that for the simplest common form of index description. (It is not always possible, for common material, to find common search procedures or output evaluation measures). A natural choice of baseline where methods of modifying given index descriptions are concerned, as opposed to providing them initially, is performance for the given descriptions, i.e. for the primary indexing, with some straightforward search procedure and output presentation.

The first possibility, choosing the simplest (or simplest common) form of indexing, most obviously applies to a given body of raw material, while the second, using the primary indexing, is natural for individual collections derived from the material. In fact, since the differences between collections as well as between bodies of material are important, for specific collections the primary indexing with simple coordination level matching has been adopted as a baseline.* For the Cranfield, Inspec and Keen data, natural baselines for the bodies of material are provided by the original manual indexing which was done in a generally, though not totally, comparable way (the main difference being indexing source). For the UKCIS material there is no obvious baseline, a problem which is discussed in more detail later.

In the Run Tables, the runs defining baselines are marked with an *.

3 Performance

3.1 Relevance Need

As mentioned in the discussion of input factors, relevance need is an important enviornmental constraint. Two specific parameters are involved, referring to the character of the individual documents retrieved,

* i.e. performance for the primary indexing of documents and main (as as opposed to alternative) requests.

and to the set of documents retrieved respectively. In the first case, which we may call quality need, we are concerned with different relevance grades. In terms of our test data characterisation, we may be interested in highly relevant documents only, or in any documents deemed relevant. In the second case, which we may call quantity need, we are concerned with whether we get some or most of the relevant documents. In conventional terms these are described as low and high recall needs; the former is often (but not necessarily) equivalent to a high precision need. Quality need may be labelled 'best' or 'any' while quantity need, given the usual properties of retrieval systems, as 'few' or many'.

None of the test data provided explicit user statements of quality or quantity need; and we are not technically competent to judge how far such needs are implicit in the initial request statements. Some of the test data nevertheless provides information needed to judge performance in relation to hypothesised quality needs, in the form of relevance grades generating relevance variants. So presumed quality needs may be considered by comparing performance for highly relevant documents with that for complete relevance sets. Thus one question is whether the relative behaviour of e.g. indexing variables is consistent under the two types of need; a second is whether, if it is not, the differences in behaviour are correlated with the needs. We have not duplicated every test for the relevance variants available, but have hopefully done enough runs to justify some conclusions on statistical classification and weighting in relation to quality need. The fact that the Cluster Hypothesis Test results show no marked collection differences for relevance variants is important here too.

Hypothesised quantity needs may also be studied using the given relevance judgements (usually the full sets). This may be done in a fairly crude and obvious way by comparing given recall and precision values and by explicitly relating them to size of output. In particular, since our tests conventionally carry out exhaustive searches and provide total output (e.g. down to coordination level 1) they may be deemed to be geared to the user wanting many relevant documents; the test results can then be processed to indicate performance for a supposedly more selective system user. Thus we may consider not only precision values for given low recall, or recall for given high precision, but both recall and precision for specific small output volumes. As will be described below, we have attempted to characterise performance in a variety of ways; and one relevant to the user wanting only a few relevant documents is Cooper's expected search length for, say, a single relevant document, which we have computed for selected runs. A sophisticated approach to quantity need would be to apply van Rijsbergen's effectiveness measure 'E' with different weightings for recall and precision respectively; but we have in fact only used this with equal weightings.

3.2 Evaluation

We have to consider two aspects of a retrieval system performance measure: its content and its form, i.e. what features of system behaviour are taken into account and how information about them is presented. We have basic evaluative notions, and specific ways in which they are applied.

In general, we have not been convinced that other notions are markedly superior to recall and precision for general, experimental, performance evaluation. (As noted above, cost-based measures are not appropriate to

projects like ours). But since recall and precision values lose detailed information about document numbers which may be important in real life, we have also considered some rudimentary but transparent characterisations of performance in terms of numbers of documents retrieved. Recall and precision information itself has been presented in different ways.

3.2.1 Measures

(a) Simple numerical performance characterisation

The actual numbers of relevant and non-relevant documents retrieved in a complete search are where appropriate listed in the Run Tables; and for selected runs with a fully ordered output, a cutoff is easily applied, the numbers of relevant documents retrieved for specific small numbers of best matching documents are given. The information about basic matching relationships given for the test collections in Figure AII.11 provides additional information relating to baseline searches.

As mentioned above, Cooper's expected search length measure (Cooper 1968), computed for a single relevant document, is given for some runs. This represents, for a request set, the average number of non-relevant documents to be inspected and discarded before the relevant document is reached, and is easily provided for a fully ordered output: in this case it is simply the average over the queries of the numbers of non-relevant documents retrieved before the first relevant documents, i.e. the number of ranks occupied. Our procedure is strictly an approximation to Cooper's measure, since replacing any partial ordering by a full ordering beforehand preempts a proper treatment of the partial ordering by the formula; but as the approach in both cases distributes equally ranked documents randomly, the effect must be much the same. The search length results given should in any case only be used to compare performance for collections derived from the same raw material with the same queries, documents and relevance judgements. Related analyses are also given showing the proportion of a request set retrieving their first relevant document by specific rank cutoffs: these are presented as cumulative frequency graphs.

(b) Recall and precision

Since there is no overridingly convincing method of characterising average retrieval performance in terms of recall and precision, we have proceeded in a simple minded way: we have used a range of procedures justified on a variety of grounds, in the hope that if different methods of indexing or searching show the same relative behaviour, our conclusions about their merits are sufficiently reliable.

As many of our retrieval methods naturally generate an ordered search output, our results are primarily presented as recall/precision graphs of the familiar kind. These take two forms. In the first case averaging by numbers across requests has been based on matching values (real or notional coordination levels). The resulting recall/precision points are conventionally linked to form a graph. To simplify plotting and aid comparisons we have applied a simple linear interpolation procedure to the given set of points to obtain precision at (up to) ten standard recall levels. (The formula is given in Appendix 1). The main results in the Run Tables are obtained in this way. It has been argued that no good theoretical claims can be made for this procedure. While not necessarily accepting this, we have adopted the method largely because it is widely used and facilitates inter-project comparisons. It is compared in some cases with similarly obtained graphs based on matching ranks, for fully and completely ordered output. An alternative view of ranked output is provided by some very

simple document cutoff figures, giving recall and precision points computed by average of numbers for specific rank positions.

As indicated in Chapter A II, the U27000P collection involves Boolean search specifications normally generating an unordered output. Recall and precision in this case has been computed by simple average of numbers.

Our second technique has been to average precision computed for individual requests at ten standard recall levels following the "pessimistic" interpolation method advocated by van Rijsbergen (1975a). This technique has been reserved for fully and completely ranked output: some matching procedures generate this naturally and other results have been forced into this form in the manner described earlier. Graphs obtained in this way are similar to the SMART Project ones, except that the latter are based on a more optimistic interpolation. Van Rijsbergen gives theoretical arguments justifying his method. We accept these, but with reserve, and have therefore produced graphs in this way mainly for results where comparisons with the SMART Project are most appropriate. (The interpolation procedure is detailed in Appendix 1).

The methods of deriving average recall and precision graphs we have adopted represent an arbitrary selection from the possibilities, and we would not wish to make any very strong claims for the theoretical solidity and rperesentative truth of any of them. Moreover, this is not the place to enter into a full discussion of evaluation. The following comments therefore simply summarise the main points involved, and indicate the type of argument which may be advanced to support the particular choices of graph generation procedure we have adopted.

We may distinguish direct and indirect representations of performance, the former obtained by averaging across observed matching values or ranks, the latter by averaging across observed performance points, or rather at standards derived from them, usually to give precision at standard recall. In the first case, sometimes called "document cutoff" (Keen 1967), whether values or ranks are selected as averaging base is an external choice depending on different views of how requests are comparable; ie comparisons between requests can be made either between documents having the same matching value or the same rank: which is preferred is an independent matter; in the second case, called "recall cutoff", the distinction is immaterial. For document cutoff there is a choice of averaging by numbers or by ratios. Unrepresented requests are simply disregarded in the former, but have to be specifically allowed for in some way in the latter. On the whole, it may be argued that average of numbers is preferable in involving a consistent treatment of given information only, though this information may be inadequate: average of ratios involves either inconsistent treatment or hypothesised values. For recall cutoff, unobserved performance points may have to be hypothesised, i.e. interpolation (or extrapolation) to cover the range of standard recall values may be needed; van Rijsbergen's argument is that his pessimistic interpolation (with dependent generating point selection), rather than optimistic techniques like those used by SMART, is the only legitimate method. Document cutoff average curves may be subsequently processed, as indicated above, to give precision values at standard recall levels: this second step is simply for convenience, and in this case simple linear interpolation is quite appropriate; extrapolation is not desirable.

The overall distinction between the two types of procedure, is that document cutoff is essentially descriptive of actual performance: it provides a picture of what happened, from which broad generalisations may be made. Thus actual performance may not exhibit a smooth inverse relationship between recall and precision, and linear interpolation relating observed points is only reflecting the transition from point to point. Recall cutoff as advocated by van Rijsbergen depends on a strict inverse relationship; it may therefore disregard some actual observed points in order to determine the performance curve underlying retrieval results for a particular retrieval technique for which the observed results are regarded as sample output. Recall cutoff may thus be deemed predictive in a strong sense, and has a very different motivation from document cutoff.

The differences in performance evaluation method just described are strictly independent of the form of the search output, though applying them to unordered output woult be futile or strange. Further, there is no reason in principle why, for evaluation by a particular method, comparisons should not be made between runs giving different forms of output, and especially without regard to whether output is partially or fully ordered. We have, however, felt it safer to confine comparisons to output of the same form. Purely practical constraints have also meant that we have not been able to compute the alternative performance representations for every run: the particular choices we have made have been determined in part by economy and in part by the need for compatibility with earlier or related work.

Thus document cutoff has been chiefly applied to partially ordered output with requests compared by matching value, but also to fully (and also completely) ordered output compared by rank. We have not compared by rank for partially ordered output, and some matching coefficients naturally generating fully ordered output, like cosine correlation, allow so many values that comparison by value is slightly ridiculous. Our choice of value based document cutoff comparisons was mainly due to a historical commitment to coordination type searching. Recall cutoff has been applied to fully and completely ordered output, following SMART practice, so partially ordered output has been forced to fully ordered. We have not applied it to partially ordered output directly, partly because the extensive interpolation typically needed in this case seems rather objectionable. It should be noted that fully ordered ranking is extremely expensive, especially for large collections, and this has limited the number of runs done in this way: indeed for the UKCIS material, only the top ranks were worked out. Most of the recall cutoff runs depend, for comparability with SMART, on the use of cosine correlation as a matching coefficient. This unfortunately implies a change in the value of a minor output variable (see Chapter IV of Section B), but this is unlikely to have much effect on the general views of comparative performance indicated by one evaluation method as against another.

An alternative treatment of recall and precision has been applied to some results. This involves the use of van Rijsbergen's effectiveness measure 'E', which is appropriate to simple unordered output: it was originally proposed for cluster based or Boolean output (see van Rijsbergen 1971 and 1975a). But ordered output can of course be treated as unordered. Given the two sets of retrieved documents, A, and relevant documents, B,

$$E = \frac{|A \Delta B|}{|A| + |B|} \qquad \text{or} \qquad 1 - \frac{2|A \cap B|}{|A| + |B|}$$

E characterises the difference between the two sets of documents, and is small when this is small. The formula may be modified to reflect a user's preference for recall or precision. In our application an equal preference is assumed. For a set of requests performance defined in terms of E may be represented as a cumulative frequency plot showing the percentage of requests achieving a given E value (so the higher the percentage for low values, the better). E figures for selected runs are given in the secondary Run Tables: in these the retrieved document set for a request is defined as that giving the best, i.e. minimum value of E over the range of set cutoffs represented by the rank positions in a fully and completely ordered output. The cutoff is variable for different requests and being calculated post hoc is not simply determined by searching. It is thus somewhat unrealistic and our performance evaluation with E is therefore rather abstract. It may be noted that although recall and precision are equally weighted, the best E values tend to be correlated with low recall.

Our evaluation procedures are summarised for reference in Figure AIII.3. It should be emphasised that as the ordering of output may be manipulated, e.g. to derive a full ordering from a partial one, or to suppress a partial or full ordering, the same searches may be evaluated and compared in different ways. Figure AIII.4 gives request numbers represented at high ranks.

In principle significance tests should be applied to confirm apparent differences of performance. It is not obvious that any standard test is especially appropriate, and for our main recall precision graphs we have therefore adopted a crude rule of thumb based on the area enclosed by the curves. We deem a difference of at least 5% is significant, and more usefully characterisable as noticeable from the point of view of a system user, while a difference of 10% is characterisable as material. (In practice the comparison is made by eye rather than program). We recognise that this approach is rather crass; but we hope that the large differences which would be the only ones likely to be of interest to users have a good chance of being genuine.

UKCIS performance

As mentioned in Paragraph A II.1, the UKCIS relevance judgements are associated only with the assessed pooled output of the original UKCIS project searches using the Boolean profiles; any documents retrieved by other indexing search methods and not assessed are therefore deemed non-relevant. This is unfair to these methods since they may in fact retrieve documents which would have been assessed as relevant if found originally. That is, if performance is evaluated only in terms of the known relevant documents for the U27000T and U27000P collections, any non-Boolean procedures may appear to perform relatively less well in relation to the Boolean ones than they in fact do. This is particularly important for any coordination type searching of our usual kind, which typically retrieves, as is evident from Figures AII.11 far more documents than the original searches.

The UKCIS project workers made various attempts to estimate real recall (see Barker 1972a). From these, and particularly their estimate of 40% real recall for title searching, we may infer that they retrieved in their pooled output about 65% of the relevant documents to be found. This information may be used to adjust our own results in various ways. Thus

for any searches involving the Boolean formulations, recall and precision can be simply recalculated in relation to a new total of relevant documents hypothesised from the UKCIS estimates. Thus if the known Boolean relevant retrieved of 6609 are deemed to represent 40% of all relevant, the hypothetical new total is 16523 documents. The Boolean search performance of 61% recall and 44% precision based on the known relevant total then becomes 39% recall and 44% precision for the new hypothetical total. For non-Boolean searches retrieving new documents, and in particular generating an ordered output, performance values may be adjusted by the following strategy. We assume that the distribution of the extra relevant documents retrieved per coordination level is the same as for the assessed, and derive new relevant and non-relevant retrieved totals for each level: these are then used, with the hypothetical total , to give new recall and precision values.

The general effect of the adjustment procedures is to revise Boolean performance downwards, and non-Boolean upwards. We have not in fact used them to characterise performance in the regular Run Tables. This is based only on the known relevant documents, and must therefore be treated with caution as not giving a true picture of absolute performance, and as probably giving a misleading picture of the size of relative differences in performance. We have instead provided some separate adjusted performance figures.