PART 1 : TEXT

Section A : Background

Section B : Experiments

Section C : Discussion

# SECTION A : BACKGROUND

## 1 Object of the work

The project research was designed to follow up initial successful experiments with relevance weighting (Robertson and Sparck Jones 1976, Sparck Jones and Bates 1977). These tests showed that retrieval performance for simple postcoordinate term requests could be substantially improved by weighting the request terms according to statistical formulae exploiting information about their occurrences in relevant and non-relevant documents. The results obtained for different weighting formulae also provided support for a probabilistic theory of indexing developed by Robertson 1976.

In general terms the intention of the project was to investigate relevance weighting in a wider range of environments and under a greater range of conditions than those of the initial experiments reported in Sparck Jones and Bates 1977. The most important specific aims were to test relevance weighting for large collections, and in searches for which rather little relevance information is available. Large scale tests and minimal information tests were required for both methodological and practical reasons: they would show whether the original performance improvements could be sustained in less favourable environments, and specifically in the kind of environments likely to be encountered in modern on-line searching. Additional groups of tests were carried out to study other aspects of relevance weighting, so as to provide comprehensive information about its properties and applications. These groups of experiments were concerned on the one hand with the comparative behaviour of different relevance weighting formulae, and and on the other with forms of query modification, for example by adding terms, and with the combination of different forms of weighting and modification. These tests included studies of both well-founded and crude formulae, and of more and less sophisticated methods of altering query composition.

The project work was conducted in the same style as that of the previous project on automatic indexing, and is reported here using the terminology and following the conventions of the automatic indexing report (Sparck Jones and Bates 1977). Some of the results given in the 1977 report are indeed reproduced for the sake of completeness, and for convenience details of some of the test data used, the methods of performance representation used, etc., are repeated here. Accounts of some of the experiments have already been published in papers dealing with particular questions about relevance weighting (see Sparck Jones 1979a,b and 1980).

## 2 Test data

The 1977 report distinguished test raw and source material on the one hand from test collection on the other. The former refer respectively to original document and query need texts and to the forms of these used for indexing, the latter to the particular index descriptions derived from these, and specifically to what was called the primary indexing descriptions. These are simple term i.e. word stem

descriptions, obtained from indexing done in manual or automatic mode. Thus a collection consists of a set of primary descriptions of documents and one of requests. These may be summarily referred to, where the intended meaning is plain, simply as documents and requests. A single body of raw material may of course supply more than one body of source material, which in turn may be utilised as the base for more than one collection. However in practice it is more usual to find a single body of raw data giving different sources, say titles or abstracts, with each source represented by only one primary indexed collection. In the previous project comparisons across different collections derived from the same raw material, and across collections, perhaps with the same type of source, derived from different raw inputs were both of interest. Both types of comparison were equally of interest in the new project.

In the earlier project automatic and manual indexing mode requests derived from a single source were (somewhat inconsistently) not regarded as generating fully distinct collections, but were merely labelled alternative requests, referred to by 'a' and 'm' respectively. In the present project there was no explicit comparison between a and m requests for effectiveness, and the labels a and m are retained only to make links with tests described in the 1977 report quite clear.

For performance evaluation relevance assessments for the queries are required, and a collection is conventionally taken as including a set of assessments. Relevance assessments are essentially part of the test context, and the previous project included some comparisons designed to see whether different sets of assessments, or relevance variants, affected the comparative performance of, say, weighted and unweighted terms. But mainly to save space in the results tables, as these tests were quite limited, they were not treated as involving different collections. The weighting schemes studied by the present project should in principle be tested in such different relevance contexts, but this was not done, largely due to the data processing effort involved. The main results in the present report therefore refer to single sets of assessments for each collection, typically including both highly and partially relevant documents. Experiments designed to study the consequences for relevance weighting performance of different amounts of relevance information were, however, part of the project work. These involved specific selections of documents from large collections, and should be regarded as representing different collection environments, though these were created artificially and did not occur naturally. The tests assumed that they were adequate simulations of real situations.

In some cases subsets of documents or requests may be supplied or created which are appropriate to particular tests. These, and particularly those used for performance evaluation, are referred to as subsidiary collections.

Two bodies of data used for the earlier experiments were used for the further tests reported here. These were one Cranfield collection, exploited partly as a convenient trial vehicle and partly because it was being used for related experiments by Harper (see Harper and van Rijsbergen 1978, Harper 1980); and two UKCIS collections, which provided the first large collections used in relevance weighting tests. To allow for full comparisons, some results for all these collections are reproduced from the 1977 report. New bodies of data for the project were obtained from the National Physical Laboratory through the courtesy of Dr

P.T.K. Vaswani, and from INSPEC through Mr L. Evans. It was regarded as essential to the project that a second large collection, or, if possible, set of related collections, be used, given the limitations of the UKCIS data. The UKCIS data supplied only titles as document source material, and two request sources, short text need statements from which term lists were derived automatically, and sophisticated Boolean SDI profiles supplying manual term lists; a more serious limitation was that relevance assessments were available only for the output of Boolean searches. It turned out, however, to be extremely difficult to obtain a 'good' new large collection: the MEDLARS data used by Barraclough et al. in their Medusa experiment (Barraclough et al. 1975) was found to provide extremely litle relevance assessment information, and it appeared that the MeSH indexing would make its use unnecessarily complicated. (It should be emphasised that to test relevance weighting based on little relevance information effectively, more comprehensive assessment information is required.) The NPL data, discovered after desparate forays into cupboards, provided document abstracts as well as titles for a large set of documents, nearly a hundred queries, and extensive assessment information. It was originally used for Vaswani and Cameron's experiments with statistical association techniques (Vaswani and Cameron 1970), and the possibility of linking newer with older research was an additional attraction. The INSPEC data was used in Evans' investigation of search strategies producing output rankings (Evans 1975a,b), and so would allow further comparisons between the relevance weighting results and those for related search techniques. In this case the set of documents was not large, and was represented only by titles, but the queries were available both as long need statements and as SDI Boolean profiles. The new bodies of data thus provided four test collections. The NPL material generated two collections with the documents indexed automatically from titles and abstracts respectively, with the same request set; the INSPEC material, called Evans to distinguish it from some quite different INSPEC material used for the 1977 tests, generated two collections each with the documents indexed automatically from titles but with requests indexed automatically from the need statements in one case and manually in the form of the profile term lists in the other.

The detailed description of the data and collections is given in Figures A1-A3. These figures give a test data summary, showing the essential features of the experimental collections; a description of the raw material referring to the original project generating it; and notes on the derivation of the collections from the source material. The details for the Cranfield and UKCIS collections are reproduced from the 1977 report, and those for the new collections are in the same style. We are grateful to Dr M.F. Porter and Dr C.J. van Rijsbergen for doing the basic processing of the NPL abstracts, as this material was very bulky. (It should be noted that minor data discrepancies occur due to processing operations and accidents: thus there were fewer NPL documents than Vaswani originally used since some duplicates were discovered and eliminated.) The new collections have been named in the same style as the earlier ones, with mnemonics referring to the original project supplier, the (rough) number of documents, and the document indexing source. Thus the NPL data generated the N11500A and N11500T collections for abstracts and tiles respectively, the Evans data the E2500T and E2500P collections for titles directly indexed and indirectly indexed via profiles respectively. For convenience the content of Figure A1 is reproduced here.

As this summary shows, the project made use altogether of seven

TEST DATA

| raw/source material name | collection name | size reqs | size docs | primary indexing |
|---|---|---|---|---|
| Cranfield | C1400I | 225 | 1400 | manual from documents |
| UKCIS | U27000T | 182 | 27361 | automatic from titles |
| | U27000P | 182 | 27361 | automatic from titles via profiles |
| NPL | N11500A | 93 | 11429 | automatic from abstracts |
| | N11500T | 93 | 11416 | automatic from titles |
| Evans | E2500T | 39 | 2542 | automatic from titles |
| | E2500P | 39 | 2542 | automatic from titles via profiles |

1) C1400I, U27000P and E2500P requests manual, others automatic;
2) many UKCIS tests with a subset of 75 requests (see below).

collections, of very different characters; taken together they represent a range of test environments for relevance weighting providing, though regrettably not comprehensively enough, both differences and similarities in the values of gross collection properties like subject, size and source. A more detailed characterisation of the test collections by numbers of terms per document and per request, by numbers of relevant document per request, and so on, is given in Figures A4 and A5. Figure A4 gives this information for the 'regular' collections, i.e. those where the indexing of both documents and requests is done explicitly via a term dictionary for the collection derived from the document set; Figure A5 gives the corresponding information for the profile collections U27000P and E2500P. For these collections there is no term dictionary, since terms are defined by requests: a particular word string or truncated word in one request may match a subset of the documents selected by a word string or truncated word in another request: document indexing is defined by matching. As the figures show, the test collections are both different from and similar to one another in a variety of ways, allowing fairly solid testing of relevance weighting. These relationships are summarily displayed in Figure A6. Thus there are 20.0 terms per document for the abstract-derived N11500A descriptions, compared with 5.7 for the title-based N11500T collection; the latter in turn has descriptions of much the same length as U2700T. Again, the E2500T and E2500P collections have very long request descriptions, with 32.4 and 48.0 terms respectively, but they are automatically provided in one case and manually in the other; the broadly comparable U27000P manually indexed requests have 29.4 terms.

Unfortunately the absence of completely systematic differences on all the major document, request, and relevance assessment properties means that comparisons designed to establish the conditions most conducive to effective relevance weighting cannot always be made. For example (disregarding the gross subject property), while there are collections combining short documents with short requests (U27000T and N11500T), short documents with long requests (E2500T), and long documents with short requests (N11500A), there is no collection combining long requests with long documents. The best that can be said is that at least some of the required alternatives are available, so some inferences for the conditions suiting relevance weighting may perhaps be made. Since the exhaustivity or, crudely, length of document and request description may be particularly important for relevance weighting, the relevant facts

for the collections are summarised for convenience below; the table also gives the average number of known relevant document: it is quite possible that the utility of relevance weighting may vary with relevance requirements, and while averages do not reflect the needs of any specific customer, they may be treated for limited experimental purposes as representing typical customers.

COLLECTION PROPERTIES

| Collection | Av. rel. per req. | Indexing | | | | | |
|---|---|---|---|---|---|---|---|
| | | Request | | | Document | | |
| | | mode | source | length | mode | source | length |
| C1400I | 7.2 | manual | need st. | short | manual | document | long |
| U27000T | 58.9 | automatic | need st. | short | automatic | title | short |
| U27000P | " | manual | profile | long | automatic | title | short |
| N11500A | 22.4 | automatic | need st. | short | automatic | abstract | long |
| N11500T | " | automatic | need st. | short | automatic | title | short |
| E2500T | 23.1 | automatic | need st. | long | automatic | title | short |
| E2500P | " | manual | profile | long | automatic | title | short |

It should be noted that though the UKCIS data was felt to be unsatisfactory through providing limited relevance information, based on assessments of documents selected by a particular type of search, the other new collections do not have exhaustive assessments like the Cranfield one. For both the NPL data and the Evans data assessments were made of search output. However this was on the pooled output from a range of alternative searches for each request, and thus has a better chance of being reasonably comprehensive. The NPL project indeed concluded after some checking that the known relevant documents probably constituted about 80% of all relevant. An additional point is that the output in these two cases came from searches of broadly the same type as those studied by the present project, unlike the UKCIS Boolean outputs, so there is much less chance of retrieving documents labelled non-relevant only because unassessed than there is in experiments with the UKCIS data. (The specific problems of the UKCIS data are discussed in the 1977 report.)

One body of data naturally generated several subsidiary collections. Thus the original UKCIS profiles were of two kinds, those with a regular Boolean structure, and those with a more complex weighted structure. The properties of the latter were so idiosyncratic that is seemed inappropriate to regard the profiles as essentially homogeneous in character; more particularly, in the previous project some tests were done both comparing and also combining the normal type of coordination strategy with the regular Boolean one, the profiles with weighted structure being quite unsuited either to comparison or combination. A subsidiary collection for the U27000P collection was therefore set up using request term lists derived only from the 75 regular Boolean profiles. This collection, named U27000Pb, was used for most of the present project experiments with UKCIS profile material. For comparison

purposes in the present project a subsidiary of the U27000T collection was set up with the corresponding 75 requests, called U27000Tb.

The particular properties of the UKCIS material meant that some other useful subsidiary collections could also be derived from it. Thus the UKCIS data includes documents taken from both CAC-1 and CAC-2, which are distinguished by subject areas within chemistry. Subsidiary collections for the two areas, representing documents numbered 1-11613 and 11614-27361 and referred to as First (CAC-1) and Last (CAC-2) respectively, were therefore set up for both the title and selected profile forms of the data, generating collections named U27000Tf and U27000Tl, and U27000Pbf and UU27000Pbl. These were used, as described below, in experiments designed to study the effect of limited subject variation on relevance weighting performance. Essential details of the properties of these subsidiary collections, together referred to as the CAC sets, are given in Figure A7.

In the original relevance weighting experiments the predictive value of relevance weighting, i.e. the value of weights derived from a search of one set of documents for searching another, was investigated via a pseudo-random division of the test collection into Even- and Odd-numbered subsets. Thus weights were calculated using information about the Even-numbered or weight generation set, and applied to the Odd-numbered or weight application set. Figure A7 gives the main details for the Even and Odd versions of the various collections, chiefly for the sake of the numbers of relevant documents involved. It should be noted that the subsets of a collection may differ slightly in both properties and performance, but these differences are not important. Finally, as mentioned earlier, various artificial collections essentially providing systematically differing relevance environments were created. Thus another set of pseudo-random subsets was created for the UKCIS collections, for experiments with weights based on systematically decreasing amounts of relevance information. In this case selecting, starting with document 1, every three out of four, every other, every fourth, every eighth and every sixteenth document gave an inclusive series of weight generation sets called Threequarters, Half, Quarter, Eighth and Sixteenth respectively. The weight application test for all of these was the document set formed from every fourth document starting from 4, which was called Search Quarter. The details of these sets, collectively referred to as the FRACTION sets, are also given in Figure A7. Unfortunately the UKCIS collection was the only one large enough to justify the considerable amount of data processing required to set up subcollections for this type of test.

Rather different subsets were created for tests studying variations only in the amount of relevance information available for weight generation. These were modifications of the Even sets, with either one, two, or three relevant documents identified in complete sets of constant size. These sets were again created in a pseudo-random manner, by taking the numerically first, first two, or first three known relevant documents (other relevant documents being treated as unknown). By extension of the 1977 report terminology these are referred to as variant collections, since they are of interest primarily from the relevance data point of view. These sets were designed to study relevance weighting based on little information in a relatively neutral manner. However since one of the project objectives was to investigate weighting in conditions like those of real systems, further weight

generation sets, i.e. variant collections, were created in a manner intended to simulate an on-line or SDI search environment. That is the one, two, or three relevant documents with the highest simple term matching scores were selected from the Even sets. (The detailed arguments for these selection strategies are given in Section B; for the present it should merely be noted that while the best matching documents might be highly relevant ones, their grades were not specifically noted.) The variant collections are labelled FIRST1, FIRST2, FIRST3, and BEST1, BEST2 and BEST3 respectively.

To complete the description of the test data, some facts about basic term retrieval are given in Figure A8. The object of the project was to improve on simple term matching, and since the results are given mainly in the more abstract form of recall and precision graphs, Figure A8 provides some supporting figures about actual numbers of documents retrieved. Note, however, that while term matching is by coordination levels, the figures about numbers of documents retrieved in Figure A8 refer to totals matched on any number of terms. The figures should therefore not be regarded as illustrating performance, but rather as indicators of the basic matching propensities of the data.

## 3 General testing strategy

The previous project work distinguished two types of retrieval system factor; these were environmental parameters not subject to explicit test control, and system variables which were controlled in the project experiments, by being assigned values. Most of the properties of the test data described above, like document subject or relevance status, were parameters for the previous project and are equally parameters for the tests described here. The properties of the primary indexing were treated as variables, though the experiments investigating them were not as systematic as could be wished. However for the purposes of the present project the primary indexing was regarded as supplying environmental parameter settings, and was not studied in its own right.

It will be clear from the account so far that as the project data supplies a range of environments for investigating the behaviour of system variables of interest, as several such variables, with different values, were examined, and as performance is represented in different ways (described below), there are a good many results to be considered.

In the 1977 report the raw retrieval figures were tabulated in an essentially neutral way, but discussed by reference to classes of system factor, and more particularly of system variable, which were used to provide an orgaisational framework for the tests. A similar approach is adopted here, though as the experiments were more concentrated, the whole is simpler.

Thus following the conventions of the 1977 report, the set of searches for a particular set of requests against a particular set of documents, with evaluation using a particular set of relevance assessments, will be called a run. A run therefore implies a particular choice of values for the various system variables explicitly studied in the project, i.e. for the primary test variables. The use of a specific collection also represents an implicit selection of settings for the global environment parameters.

The primary variables studied in the earlier project were characterised in the 1977 report under the headings of input, indexing, and output factors. Input variable choices included, for example, description length, indexing variables included, for example, term classification, and output variables included, for example, matching condition. The choices of variable value, being made for a particular collection, were of course related to the primary indexing descriptions of this collection. As noted, the experiments described here were entirely focussed on indexing factors: input factors were regarded as parameters and thus studied only indirectly and informally via comparisons across collections. Thus, referring to the input factors listed in the 1977 report, the various collections covered differences in indexing mode, indexing source, indexing description exhaustivity, though not, explicitly, indexing vocabulary specificity. The output factors investigated in the previous project were scanning strategy, matching condition, and scoring criterion. In the present project some very limited comparisons between normal coordination level matching and Boolean searches were made, which refer to matching conditions, and the standard comparison between weighted and unweighted terms refers to scoring criteria. However the reference to matching conditions was so marginal that it cannot be regarded as a serious treatment of an output variable. Further, there were only two scoring criteria, and their use was such a natural consequence of the choice of indexing variable values, that there is little point in considering them as constituting output factor studies in their own right.

The indexing factors investigated are discussed in detail in Section B: for reference the input and output variables involved in the tests in a subordinate role are summarily listed below.

SUBORDINATE TEST VARIABLES

| | |
|---|---|
| Input: | indexing mode |
| | indexing source |
| | indexing description exhaustivity |
| Output: | matching condition |
| | scoring criterion |

The results for all the runs done are given in the tables in Part 2. It will be evident that an individual run can be looked at from different points of view, and can figure in different comparisons emphasising different variables. Any particular comparison will thus focus on a specific test variable, the values of the other primary variables in particular being in this case treated as secondary. The general strategy adopted in considering the earlier studies is thus applied here: the results for each comparison with respect to some test variable and given set of secondary variable values are cross checked by the same comparison with a change in at least one secondary variable value. This should provide some solid ground for drawing general conclusions about the test variable in question.

In the 1977 report an attempt was made to ensure that the cross checks for a variable in one factor class included changes in at least

one other class, rather than to a variable in the same class. The concentration in the present project on one class only means that this is not possible and cross checks essentially utilise the different groups of indexing variables studied as described in Section B. The object of having different collections and, more importantly, different bodies of raw material was of course to test for the effects of global environment parameters; so wherever possible, though the data did not always allow or justify it, comparisons relative to the behaviour of the test variables were made across collections. In this connection, it should be noted that as the main object of the entire project was to confirm the superiority of relevance weighted to unweighted terms, within the broader context of studying methods of improving simple term matching performance, the runs representing basic term matching for all the search collections provide what was called baseline performance in the 1977 report.

## 4 Performance representation

In the 1977 report a whole range of performance representation techniques was used in an attempt to provide a firm foundation for general statements about the relative merits of different indexing techniques. As it continues to be the case that no one method, embodying both a particular measure and particular request averaging technique is unequivocally acceptable, (see Sparck Jones 1977), results are again presented in several different ways.

Most of these methods of performance representation utilise recall and precision as performance measures, but vary in the method of averaging. The chief method of representation used, named dv (document value) averages by numbers across requests on matching values (e.g. coordination levels); linear interpolation is then used to obtain precision values for 10 standard recall values. The method is one type of document cutoff technique. However since this method, though used, for example, in the Cranfield tests (without the interpolation), has been criticised, some results are presented using the main alternative method, involving recall cutoff. This method, named rc (recall cutoff), averages over the precision values obtained for each request at standard recall by pessimistic interpolation. and is the technique ordinarily used by the SMART Project. Unfortunately the recall cutoff method is very expensive, so only a few sets of results are represented by it: these are hopefully sufficient to support the conclusions based primarily on the document value results.

In general terms, the type of matching investigated by the project results in an ordered search output for a request. Boolean searches produce, for documents with a positive match, unordered output. Some types of performance evaluation require simple retrieved sets, like those produced by Boolean searches, and these can of course be obtained by applying a cutoff to an ordered output. However the essential objective of approaches to matching like those embodied in relevance weighting is that they order search output, specifically with the intention of placing relevant documents above non-relevant ones. Thus though some simple performance characterisations based on sets may be useful, the main evaluation procedures used by the project depend on ordered output. However the methods differ according to whether they are applied to partially ordered output, which may have more than one

document with the same score, i.e. rank, or to fully ordered output, with only one document per rank, perhaps achieved by forcing an ordering on documents with the same score. The document value method is applied to partially ordered output, the recall cutoff method to fully ordered. Some methods, moreover, like recall cutoff, require a complete ordering of a collection, i.e. assume that every document is ranked. It should be noted that the natural output form of a search may be transformed for evaluation purposes into another output type: an example is forcing a complete ranking.

The rankings supplied for recall cutoff can fortunately be utilised for some other methods of presenting results, so these have been used for the relevant searches. They are two more document cutoff methods, averaging by numbers across matching ranks, called dr (document rank), and averaging by numbers for precision and recall at specific ranks, called pr (precision rank).

In addition some simple numerical methods of performance characterisation have been used. All the runs are described by average number of documents and relevant documents retrieved with positive matching values: this is the type of information given in Figure A8 for the primary indexing descriptions, and is perhaps rather marginal as a method of performance representation, though it supplies interesting data. It is referred to as tr (total retrieved). The other numerical methods use the completely ranked search output. Average numbers of relevant retrieved by specified high rank positions are given as rr (relevant rank). In addition, though this is not narrowly a form of performance representation, cr (cumulative requests) gives the cumulative proportion of requests retrieving their first relevant documents by specified rank positions.

The various methods used in the earlier project are fully described in the 1977 report; the details for the methods used in the present project are reproduced in Figure A9. For convenience, the methods are listed briefly here.

PERFORMANCE REPRESENTATION

| | |
|---|---|
| A. Document value (dv) | : recall/precision, matching values |
| B. Recall cutoff (rc) | : averaged precision, ranked output |
| Document rank (dr) | : as dv, but matching ranks |
| Precision rank (pr) | : recall/precision, specific ranks |
| C. Total retrieved (tr) | : average retrieved, lowest value |
| Relevant rank (rr) | : average retrieved, specific ranks |
| Cumulative requests (cr) | : requests retrieving, specific ranks |

It should be noted that, as described below, the organisation of the tables containing the search results is in part by performance methods: the main tables give performance characterised by the document value method, which is provided for all sets of searches, while the secondary tables give alternative characterisations of performance using the other methods.

## 5 Tabulation of search output

The presentation of the run figures is in the style of the 1977 report.

Clearly, as individual runs can be placed in many different contexts, they can only be categorised in the Part 2 search output tables in a very simple way. Specifically, while individual runs can be placed in different contexts, to avoid repetition they have to be categorised in one way only in the tables. The tables therefore group runs in a fairly straightforward way corresponding to the main groups of indexing factor treatments discussed in Section B. Each table refers to all the search collections in relation to one type of relevance environment. Thus the tables relate primarily to the original relevance assessment data for the collections, and there are alternative tables for the variant collections involving different relevance assessment data. The parallel set of runs for different collections is called a run set. Each table therefore covers a series of run sets of a related character, the member sets being arbitrarily numbered. Altogether there are 12 tables categorising runs for the original relevance assessments as follows:

RUN TABLES

| | |
|---|---|
| T | Terms |
| W | Weights |
| R | Relevance weights |
| S | Substitutions |
| SW | Substitutions with weights |
| SR | Substitutions with relevance weights |
| C | Classifications |
| CW | Classifications with weights |
| CR | Classifications with relevance weights |
| E | Expansions |
| EW | Expansions with weights |
| ER | Expansions with relevance weights |

As noted, the explanation for these table descriptions is provided in Section B.

Runs involving the systematically related variant collections are grouped in separate tables, tagged v1/ for the Sixteenth, Eighth, Quarter, etc. set of FRACTION variants, and v2/ for the FIRST1 etc. and BEST1 etc. sets of variants; the individual run sets in these tables have the same names as those in the regular collection tables. (The UKCIS CAC collection involved only one weight generation set, so the results are given in the main tables.) Further, as several different performance representation methods were used, there are sets of tables for the same collection searches but with performance shown in different ways. The notion of run in fact includes the method of performance representation used, so the same searches represented in, say, two different ways will be referred to as two runs. As indicated earlier, one method of performance representation, namely by document value, has been used as

the main method of representation. The tables for these runs are therefore referred to as Main Tables, while those for the other methods used a referred to as Secondary Tables, being distinguished by the specific method used. Thus any Secondary Table tagged with 'rc' refers to recall cutoff representations. The corresponding run sets in Main and Secondary Tables are always given the same identifying number. As mentioned earlier, the secondary representation methods were quite expensive, so the Secondary Tables are much less complete than the main ones and serve chiefly as controls.

The tables are preceded by a Key Table giving the run set characterisations. This is followed by a Summary Table indicating what runs have actually been done for the different collections, and so providing a lead into the detailed tables. The Summary Table naturally refers to all the outputs appearing in the Main Table, and also the tr Secondary Table; the outputs with corresponding entries in the other Secondary Tables are marked by '.

The tables are used to provide a systematic way of naming runs. Thus references to runs in the text of the next section presenting the project results will have the form MX9 or SrcX9, say, referring to run set 9 in group X, and Main Table document value and Secondary recall cutoff representations respectively. The corresponding variant collection run set names have the form v2/MX9 and v2/SrcX9. A series of runs in the same table, say MX1, MX2 and MX3, will be referred to as MX1-3. Individual runs in a set are identified by their collection name, for example the run for E2500T in set MX9. If M and S are omitted, all the corresponding tables X9 are referenced. Thus an important run set is that labelled T1, which gives baseline performance for all the collections, for the different methods used.

Though the search outputs are primarily given in the tables, some particularly interesting comparisons are given as recall/precision graphs of the usual kind.

Finally, there are some miscellaneous figures not fitting the standard scheme: these are given in Other Table.

It is impossible to print all the test results in the form of conventional recall/precision graphs: however as graphs are more pleasant to study than columns of figures, the more interesting test results are illustrated by Graphs in Part 2. These graphs are chiefly for the main document value performance representation method, but some alternative recall cutoff graphs are also given (for convenience those stricly paralleling the main graphs are given corresponding page positions).

## SECTION B : EXPERIMENTS

In the previous section the object of the work reported here was briefly introduced as testing relevance weighting. The general approach to testing was then presented, utilising the ideas of system factor and variable value comparison. The project work was concerned with indexing factors, and specifically request term relevance weighting, and hence with experiments comparing weighting variable values.

This section is devoted to the substance of the project tests. The variables studied are discussed, and the value comparisons and cross checks on these presented in detail. The results are more broadly evaluated in Section C.

For these purposes some terminology introduced in the 1977 report is useful. As indicated, an individual run represents particular parameter settings and variable values. Insofar as a run involves at least one explicit assignment of a value to a variable we could describe it as constituting an experiment rather than an investigation; however it is more helpful to view experiments as comparative. An experiment therefore strictly compares two or more runs differing in the values assigned to some specific variable. Cross checking as described earlier is an essential constituent of experimentation; and in the retrieval system context it is proper to include in an experiment the comparison and its cross checks across different collections, to allow for the influence of parameter settings. An experiment thus refers to sets of runs each on as many collections as are available, i.e. to run sets for different primary and also different secondary variable values. By extension, we can naturally refer to a group of experiments on related variables.

The choice of experiments is motivated by what were referred to in the 1977 report as topics, with the experiments designed to validate some propositions relating to these topics. At the highest level the topic of this report is relevance weighting, so the experiments described are intended to validate the proposition that relevance weighting is a good thing. However this global topic is more usefully dealt with as a collection of topics referring to indexing factors, as elaborated in this Section; the experiments described below were therefore designed to validate a number of specific propositions about the form of, and conditions for, relevance weighting. These propositions included, for example,

1) that relevance weighting is valuable for different subject areas,
2) that relevance weighting is valuable in different relevance conditions,

and

3) that relevance weighting is valuable with different primary indexing inputs,

and a variety of other, more detailed ones.

The individual propositions underlying the tests are most conveniently introduced through the account of the experiments which follows. They will be explicitly listed in the Conclusion to Section C, where the extent to which they have been validated by the project work

will be considered.

## 1 Overview of the experiments

The project experiments are related to one another within an overall scheme. This is presented below first through a simple categorisation of the tests done, and then by discussion of the basis for the categorisation.

Essentially the project was concerned with the investigation of different criteria

a) for including terms in request term lists; and
b) for weighting terms in request term lists.

### a) request membership

The studies relating to request membership were concerned on the one hand with request replacement, i.e. the substitution for a given term list of another set of terms, and with request enlargement, i.e. the addition of other terms to the given list. The approaches adopted were relevance oriented, i.e. the tests dealt primarily with the inclusion in requests of terms known to be in relevant documents. Thus they involved both substituting lists of relevant document terms for original request lists, and adding relevant document terms to requests. These methods therefore fall into the general category of relevance feedback techniques studied by the SMART Project more than a decade ago (see Salton 1971).

The approaches are all defined for the purposes of the report as being concerned with indexing rather than input factors, though substituting relevant document terms for the given request terms is really an input rather than indexing factor operation. The tests with this form of request formulation do not, however, have any close links with those reported for input factors in 1977, and (illustrating again the ambiguity of factor definition in information retrieval) they are much more closely linked with the other indexing factor tests reported here.

The original motive for the experiments was on the one hand to see how the various ways of using relevance information perform when treated as alternatives to weighting, and on the other to see how effectively they can be combined with weighting. However as will become apparent, it seems to be an error of principle to regard relevance inclusion and weighting strategies as competing, and the main reason for distinguishing them is the purely experimental one of studying individual variables.

A few tests were concerned with the inclusion in requests of terms presumed rather than known to be relevant. The methods used were classificatory ones, and the tests done were primarily intended to link the project experiments to those being carried out at the same time by Harper and van Rijsbergen (see Harper and van Rijsbergen 1978), since their approach to classification, unlike those reported in Sparck Jones 1971 and in Sparck Jones and Bates, ultimately ties classification explicitly rather than implicitly to relevance.

Classification- or association-based strategies for alte.ing request term lists by explicitly using relevance information are much more sophisticated approaches to relevance feedback than the strategies described above. However, though association strategies of this sort have been proposed by van Rijsbergen 1977, as they are very difficult to implement, as will be discussed in more detail later, classification strategies not directly using relevance information may be exploited instead. The classificatory approaches to request membership tested may therefore be regarded, in relation to the choice of terms, independent of any later weighting, as making much weaker use of relevance information than the feedback approaches, but as nevertheless tacitly relying on it through the tendencies of request terms and their close associates to cooccur in relevant documents. It will be evident that with classifications not explicitly using relevance information, replacing original request term lists by new ones has little justification. The classificatory strategies studied were therefore used only to add terms to requests.

b) request weighting

The approaches to weighting investigated were also relevance oriented and, as in the 1977 report, are clearly concerned with indexing factors.

The tests involved different forms of relevance weighting, chiefly using formulae studied in the earlier report, but also some others. They were all of the type exploiting term occurrences in relevant (and non-relevant) documents. One weighting procedure not exploiting relevance information, namely collection frequency weighting (the SMART Project's inverse document frequency weighting) was also tested, partly for continuity with earlier work, but chiefly for comparison with relevance weighting. Thus the use of collection weighting, like that of collection classifications, relies on the presumed rather than given relevance behaviour of terms.

Relevance weighting is clearly a form of relevance feedback, though the type of weight used does not seem to have been envisaged, even in the abstract, in such early discussions of feedback as Rocchio's (see Rocchio 1971). A broad view of relevance feedback including relevance weighting is adopted here, following the suggestion of van Rijsbergen 1979. Thus the project tests have all been concerned with the effectiveness of different relevance feedback techniques.

To summarise, the tests described in this chapter are concerned with two indexing factors, namely request term membership and request term weighting. At a lower level under the first we have two types of strategy for obtaining term lists, namely by substitution of other terms for the given terms, and by addition of other terms. As replacement not relying on relevance information was not studied, the only replacement option investigated was relevance substitution; however for addition, enlargement without and with explicit relevance information were both investigated. The choices at the level below these which were actually tested are described in the detailed account of the experiments given below. For weights the situation is simpler, with an initial choice of weighting without and with relevance information, and with the single possibility of collection occurrence-based weighting under the former and of relevant document occurrence-based weighting under the latter. The

specific formulae tested are given later.

We therefore get a list of the indexing factor options studied as follows:

membership
- replacement
  - relevance based : substitution
- enlargement
  - collection based : classification
  - relevance based : expansion

weighting
- collection based
- relevance based

Overall, the indexing possibilities investigated by the project can be seen as aimed at progressive improvements of simple term matching performance. Thus the crudest indexing and retrieval technique is simply to take the given request terms and do coordination searching: referring to the organisation of the search result tables introduced in Section A, this gives the baseline performance of the T tables. Then weighting based on rather weak information, namely about the collection distribution of terms, gives us the W tables. Weighting on the rather richer base of relevance as well as collection information, again applied to the initial query terms, gives us tables R. Within the general framework these three tables as it were exhaust the options for requests confined to the original query terms. The next possibilities are those offered by substituting other term lists for the given lists. The natural ways of doing this would effectively parallel the contrast between implicit and explicit use of relevance information embodied in the use of collection or relevance weights: i.e. they would exploit statistical associations and relevant document terms respectively. However it is not very meaningful to consider term associations without any reference to request terms, so there is no table A. It is more fruitful to consider the use of relevant document term lists as substitute requests, giving table S. The terms in such lists can of course be collection or relevance weighted, giving tables SW and SR.

Another range of possibilities is presented by the idea of enlarging the given request terms lists with additional terms, the latter being obtained by either of the two approaches just considered, namely by the use of class-related, or associated, terms, or by expansion through relevant documents. This gives tables C (effectively meaning T+A), and E (i.e. T+S), and as the terms in either case can be collection or relevance weighted, we get CW and CR, and EW and ER, respectively.

Each of the possibilities just listed in principle subsumes many lower level options: for example the exact way in which relevant document terms are used to expand requests, or the way in which information about the distribution of terms in relevant documents is exploited in weighting formulae. However the project could not attempt to investigate large numbers of these. The methods studied were therefore selected for one or the other of two reasons. In some cases, notably the choice of relevance weighting formulae, the techniques studied were justified by both theoretical arguments and past performance, and the object of the project was to obtain further experimental support for them. The collection weighting technique studied was also selected by past performance.

The various methods of altering request term list membership using relevance information, on the other hand, were selected chiefly as natural and obvious ones, to some extent representing actual user behaviour; there is also a connection with some tentative experiments reported in Sparck Jones and Bates. However as past experience suggests that obvious strategies are not always well-founded, the choices made must be regarded as exploratory rather than properly motivated.

One of the classification-based strategies is, however, as indicated earlier, more strongly motivated, and in fact provides the clue to the underlying rationale for the project choices of types of indexing strategy to be investigated.

## 2 The framework for the experiments

Earlier work on collection-based classification, as reported in Sparck Jones and Bates, for example, was not very successful. But no attempt was made then to combine such classificatory ideas with the explicit use of relevance information. The work on relevance weighting following Robertson and Sparck Jones 1976, on the other hand, not merely made no reference to term associations: the formulae studied assumed term independence, and so precluded any motivated way of using associations. However this was recognised at the time to be a simplification of the real situation in retrieval systems, where term dependencies are likely to occur, and van Rijsbergen's 1977 dependence model explicitly makes the theoretical connection between classification and weighting. In this model, classification is based on distributional information about the cooccurrences of terms in relevant and non-relevant documents, and the theory leads both to the choice of terms for enlarging a query and to the provision of weights for the original and new terms in a request. This model, in other words, integrates the retrieval system elememts, request membership and request weighting, treated more simplistically as separate in the project reported here.

But if no relevance data is available, classification can only be collection based, and the same will be true of weighting. Further, if terms are not in fact dependent on one another, or do not in fact have different collection frequencies, classification and weighting will be of no utility. We can thus envisage a range of situations in which less information is available than in the ideal case, but for which coherently related indexing/searching strategies are suggested, deriving from the optimal one.

The elements of such a scheme are in fact available, and the scheme itself underlies the project work, though it does this, for reasons which will become apparent, in a rather informal way.

The scheme relating indexing and searching strategies in different situations is as follows. We have two aspects of requests to consider, their term membership and their term weighting, and two types of information about terms, namely collection data and relevance data. Relevance data clearly takes precedence over collection data if the former is available. Van Rijsbergen's theory says that if relevance data is available, it should be exploited to provide both additional terms via the classificatory structures defined by maximum spanning trees (MSTs) and expected mutual information measures (EMIMs), and to provide term

weights of the probabilistic type used the the previous and present projects, but calculated using term set rather than single term information. Further, if dependence information is not available (i.e. terms are not dependent on one another), but relevance information is available, weights defined by formula F4 in Robertson and Sparck Jones 1976 are appropriate, since these are what the model reduces to in this special case. If there is no dependence information, and also no relevance information, a weighting possibility nevertheless remains, namely using collection frequencies. Croft and Harper 1979 have moreover shown that, given only information about term frequencies, on certain assumptions F4 weights (approximately) reduce to the collection weighting formula used by the project, which may, following the 1977 report, be labelled F0. The remaining possibility is where there is dependence information but no relevance information: the correct strategy in this case would be to combine the MST classification with collection frequency weights. This has been tried by the project.

The worst situation is where the collection information does not in fact vary for (search) terms, so collection frequency weighting reduces to no weighting, i.e. no weights distinguishing one term from another. However, the many tests conducted with collection frequency weights suggest that small performance gains can usually be made with collection weighting, so it is perhaps more sensible to take the use of query terms so weighted as the starting strategy, rather than the given query terms without any weights at all.

It is of course possible that the theory just outlined is mistaken in that while it is right to seek to combine collection and relevance data in finding terns for request lists and in giving them values, the specific classificatory and weighting formulae advocated by van Rijsbergen are wrong. There is, however, a good deal of support for the general probabilistic approach involved.

A more important difficulty, pointed out by van Rijsbergen himself, is that there are many problems in applying the theory. These are not only practical ones arising from the complexity of the operations involved, but intractable ones of theoretical detail, especially in the area of estimation. The approach adopted by Harper and van Rijsbergen in experiments has therefore been that of seeking simpler procedures which can be viewed as approximations to those required by the theory. One possibility is to combine the use of an MST classification constructed without explicit reference to relevance information with relevance weighting of the redsulting enlarged requests; another is enlarging requests with relevant document terms subject to dependency checking using EMIMs.

The first possibility was one of the classification strategies tested by the project. But as even this type of procedure is quite effortful, the other project strategies for altering request membership by using relevant documents can be seen as further, cruder approximations motivated by cost considerations.

The request list alteration methods investigated by the project thus have only a rather weak theoretical justification; i.e. the most that could be said for them, if they proved effective, would be that they represented useful approximations. Van Rijsbergen's theory nevertheless provides a motivation for them, though it should be said that this was

supplied only during the project and not at its inception: the original rationale for the relevance substitution and expansion procedures was that they seemed an obvious kind of thing to try.

The project's main concern with relevance weighting was, on the other hand, thoroughly supported by theoretical argument, especially for weights calculated using F4, as well as by the practical interest attaching to those elements of the whole theoretical scheme which are cheapest to apply. The main incentive for studying some other relevance weighting formulae was therefore not an economic one: the object was to provide more support for the theory.

The relationship between the project work and other research, briefly indicated here, will be considered further in Section C, when the test results have been presented.

## 3 Relevance weighting

As mentioned earlier, the concern of the project as a whole with indexing factors means that apart from the general validation provided by applying the same procedures across different collections, more specific cross checks for tests on membership variables have to be supplied by altering weighting variable values, and vice versa. However, since the tests on membership were fairly limited for reasons which will become apparent, the cross checks for membership have relied only on the one form of collection frequency weighting studied, and on one form of relevance weighting, namely that using formula F4. More importantly, cross checks on membership could not be provided for many of the weighting comparisons, and the main checking for these has been supplied by the use of several different collections.

The most convenient way of organising the material of the rest of Section B is therefore to preface the account of the tests, which deals first with the membership tests and then with the weighting tests, with a brief description of the collection frequency and F4 relevance weighting formulae referred to in the membership as well as the weighting tests. Both have already been discussed in detail (see Robertson and Sparck Jones 1976 and Sparck Jones and Bates), so there is no need to consider them at length. However, since relevance weighting has been the focus of the project, the presentation is self-contained, and covers other formulae closely related to F4 which were studied by the previous project. Placing this summary before the detailed discussion of the tests also emphasises the central role of relevance weighting in the work; the term membership experiments were subordinate to the weighting tests, and were to a considerable extent intended to throw light on the value and use of weighting.

For a specific term we define
n = the number of documents in which the term occurs; and
N = the number of documents in the collection.
For collection frequency weighting, we relate n and N and, since requests generally have rather few relevant documents, we do this to assign higher weights to rarer terms. The formula for this introduced in Sparck Jones 1972 was

$$w = -\log(n/N);$$

but, as explained in Sparck Jones and Bates, F0 has been implemented as

$$w = -\log(n/\max n)$$

where max n is the highest term frequency in the collection.

Clearly, with such a weighting scheme, the weight for a term is the same for all the requests in which it occurs. However if we now take relevance information into account, we consider the weight of a term in relation to a specific request. Thus for a term in a given request we define

r = the number of documents relevant to the request in which the term occurs; and

R = the number of documents relevant to the request.

For each term in a request we then have a contingency table giving the distribution of the term in relevant and non-relevant documents:

TERM DISTRIBUTION

| Document indexing \ Document relevance | + | - | |
|---|---|---|---|
| + | r | n-r | n |
| - | R-r | N-n-R+r | N-n |
| | R | N-R | N |

Different weighting formulae may be constructed by selecting different elements from this table. Such formulae will relate the relevance occurrences of a term to all its occurrences, or more specifically, to its non-relevant occurrences. One simple possibility originally studied by Miller 1970, and labelled F1 in Robertson and Sparck Jones 1976 is

$$w = \log((r/R)/(n/N)).$$

This relates relevance frequency to total frequency, but Robertson and Sparck Jones argued in favour of formulae relating relevance frequency to non-relevance frequency, and specifically for formula F4, namely

$$w = \log((r/(R-r))/((n-r)/(N-n-R+r))).$$

The various options for using the contingency table elements can be defined in terms of alternative views of retrieval system data, and of alternative selections of retrieval output ordering. Thus the data may be described by one or the other of two Independence Assumptions, namely

I1 : that terms are distributed independently in relevant documents, and in all documents; or

I2 : that terms are distributed independently in relevant documents, and in non-relevant documents.

The search output ordering may be based on one or the other of two Ordering Principles, namely

O1 : that matching depends on the presence of request terms in documents; or

O2 : that matching depends on both the presence and absence of request terms in documents.

In principle there are four possible combinations of Assumptions and Principles, generating the four weighting formulae F1 - F4 introduced in Robertson and Sparck Jones, as follows:

However it was argued in Robertson and Sparck Jones that F4 should be

FORMULAE RATIONALES

| Ordering principles \ Independence assumptions | I1 | I2 |
|---|---|---|
| O1 | F1 | F2 |
| O2 | F3 | F4 |

selected. This is because Assumption I1 is internally contradictory, so I2 is the only proper assumption to make about the data, while Principle O2 is preferable to O1 in that it uses more information and should therefore give a more discriminating result. Certainly, in the previous project tests, F4 was superior to the other three formulae.

## 4 The tests

The tests are divided into two main groups, each with two subdivisions. The main groups are those dealing respectively with the request membership indexing factor and the request weighting factor. Each is subdivided according to whether the tests were done within the environments represented by the regular collections, or those represented by the variant collections defining different relevance conditions (including the subsidiary UKCIS collections defined by subject). The tests with variant collections relating to request membership are very limited, but those for weighting are much more extensive and important.

The descriptions of the tests in each group follow a standard pattern. The various options studied are introduced, and the test results at the appropriate different levels of comparison are presented using the main method of performance representation introduced in Section A. These comparisons are of a narrow descriptive kind: the results are more broadly evaluated in Section C. As noted in Sparck Jones and Bates, it is often extremely difficult to make any generalisations over a range of search results, and the remarks made are necessarily of a simplifying kind. Following earlier conventions, two individual sets of recall and precision values, as drawn on a standard recall/precision graph, are described as noticeably different if there is an area difference of more than 5% between them, and materially different if there is a difference of at least 10%. The symbols used, for two runs A and B, where A and B are the same (i.e. are not even noticeably different), where A is noticeably better than B, and where A is materially better than B, are respectively A=B, A>B, and A>>B. If there is a somewhat variable relationship, or different relationship along the R/P curve, we may have, e.g. A<=B. Further, if A and B on the one hand are compared with C on the other, we may have A/B>C, say. It is of course possible that observed differences are not statistically significant, but we feel that material or even larger differences, for large collections, are likely to be genuine. However the absence of significance tests on the results must be deemed a defect of the project work, only partly excusable by the fact that it is not at all obvious what tests to apply, and that the difficulties of generalising over sets of significant but heterogeneous results remain.

The form of the generalisations made is the same as that for two runs; in some cases the generalisation refers to comparisons at the most specific variable value level, over different collections; in others the generalisation is more summary, referring to higher levels of variable value choice. In some cases qualification is necessary, but in general the symbolic characterisation is only used when all the experiments concerned support it.

As the secondary methods of performance representation were less widely used than the main ones, the conclusions to be drawn from them are considered at the end of the environment divisions for each factor.

## 4.1 Request membership

### 4.1.1 Regular environments

The relevant document-based approaches are most conveniently considered first.

As indicated earlier, the original intention in studying different ways of exploiting relevance information to provide or modify requests was on the one hand to compare this way of using relevance information with that represented by weighting, and on the other to see whether the two ways of exploiting relevance information for identifying and weighting terms could be effectively combined. However it became apparent that combination is in principle the only proper approach, since relevance information is being used for two different purposes, so the tests using relevance information only in relation to request membership should be regarded as a means of determining the respective contributions of the membership and weighting components to the combined strategy. In this section the different ways of altering request membership are nevertheless considered with reference to weighting only as a possible influence on the performance obtained with different kinds of request list: i.e. the weighting runs are treated as cross checks on the membership tests. Performance for combined strategies is considered primarily under weighting given that, as indicated earlier, the membership procedures are all rather crude treatments of this indexing factor, compared with those adopted for weighting. The combined strategies are thus more naturally treated under their more fully-investigated and exigently-approached component.

It has sometimes been argued that a request in the form of a term list taken from a relevant document (or from a set of such documents) is superior to an a priori list of terms of the usual kind, so an initial choice relative to request membership is between request term lists as ordinarily supplied and terms from relevant documents. However it has also been argued that if a request is supplied in any other form than as a relevant document, the terms in the original request have a special status: they were chosen by the user. Thus while it might appear, formally, that the origins of the terms in a request list are irrelevant and that a list should be treated as freely modifiable in a succession of feedback operations, it is perhaps desirable to flag original terms and never permit them to be removed in automatic feedback modification.

These alternatives were investigated by the project, in a rather simple way: the original term lists were compared with substitute lists taken from relevant documents, and with lists including terms from

relevant documents. The comparison between the three alternatives should throw light on the respective contributions and joint effectiveness of terms from the two sources.

Below this initial set of choices between original term lists, substitute lists, and enlarged lists, there are clearly various specific possibilities for obtaining relevant document terms for requests, even when only very simple methods are considered. One question is whether the number of relevant documents used matters. Another is whether the use of 'random' relevant documents as opposed to selected ones has much effect on performance. It would be possible, for a collection with a high average number of relevant documents per request, to take one, two, three, four ... n relevant documents to provide terms for requests, and further, having selected these documents, to use them for searching the remainder of the collection. However, as will be discussed in more detail in the section on weighting, it was found more satisfactory in studying relevance weighting to work with the Even/Odd collection subsets for weight generation and application, and as the request membership tests were to be related to the weighting ones, relevant documents were taken from the Even subset for use in searching the Odd. An additional reason for doing this was that in tests with small collections, using the best-matching relevant to provide terms for requests implies evaluating performance for them with a few, less well-matching relevant documents. As in the weighting experiments, it was assumed that Even and Odd sets were like one another, so the results would be like those obtainable either from searching successive similar collections in an SDI situation, or in iterative searching of a large collection with a good many relevant documents, including both highly-matching matching documents not already inspected and less well-matching ones.

Tests with the different techniques for altering request term lists were in fact, for reasons to be discussed below, only carried out with the Cranfield C1400I collection, and specifically by searching on the C1400Io collection. The Even subset of this collection averages only 3.7 relevant documents per request, so the investigations of numbers of relevant documents to be exploited made use of one, two, or all relevant documents. (There seemed little point is trying three as well.) The one or two relevant documents were either the numerically first in the supplied relevance sets, which were deemed to be random relevant documents, or the best matching in the simple term search of the Even set. The two alternatives could be taken to represent on the one hand the user's prior knowledge of germane documents, and on the other the situation in an SDI or iterative on-line searching environment. The various possibilities are labelled 1, 2 or ALL for the number of relevant documents used, and F or B for the first or best matching choice.

As noted, these uses of relevant documents were viewed primarily as rather simple but obvious strategies which were worth trying. Rather more specifically, in relation to the overall scheme for exploiting different types of information, they were regarded as very crude ways of approximating the use of term dependencies: terms cooccurring in relevant documents are prima facie dependent on one another. However the initial tests done with the C1400I collection were not especially exciting, and as it appeared that the use of relevance weights was much more profitable, the tests were not repeated with other collections. The general Cambridge rule of thumb has been that anything which does not work on the Cranfield data is most unlikely to work on anything else,

though the reverse is not true. It is, however, possible that collections with rather different properties would benefit from the more exhaustive requests likely to follow from the use of relevant document terms. Thus as past tests on description exhaustivity suggest, short requests might be profitably supplanted by longer ones if document descriptions are short. Indeed the fact that, as will be seen, request expansion in the most favourable conditions was of use even for the C1400I collection suggests that further tests should have been carried out.

The fact that they were not was due in part to a failure to see what the test results really showed at the time when they were carried out, but more importantly to problems which appeared when the attempt was made to investigate the substitution/expansion strategies for the other collections available in the earlier stages of the project. The U27000T collection in general performed so badly, or at any rate appeared to perform so badly for the reasons connected with the relevance assessments discussed in the 1977 report, that it was felt that the effort of conducting tests with it would not be justified.

However a much more interesting problem was encountered with the U27000P and E2500P profile collections then available. This is the question of how, in request expansion, the existing and new items should be related to one another for weighting purposes. In general in such collections requests and documents are not indexed by the same type of entity. In collections like C1400I, both requests and documents are indexed via a common stem dictionary. Thus any term occurring in a document will either be identical with a request term or quite distinct from it. In profile collections on the other hand, requests are indexed by user-defined terms which may be fragments (representing real or pseudo stems), words, or word strings. Documents, in contrast, are indexed by words, or perhaps lexically-motivated stems. Request and document terms may therefore be neither identical nor distinct, but may overlap; and the question then is how, when an existing and new term overlap, weights should be assigned. If they are treated as identical, one has to be artificially subsumed under the other, thus falsifying the real distributional status of one. If they are treated as distinct, their common concept will be doubly weighted. It is possible that the best strategy is the crude one of weighting existing and new items quite separately, but this was felt to be rather unsatisfactory, and as the processing involved in profile expansion was substantial, the idea of tests with the profile collections was abandoned. The tests reported here were therefore confined to the C1400I collection.

The tests are best considered from the bottom up, and will be described first for the main method of performance representation. Thus we consider first the 1/2/ALL options for substitution and expansion respectively, regardless of whether the relevant documents are F or B. Run sets MS1-5 and ME1-5 show that for simple term matching, there are no differences, i.e. that 1=2=ALL. Further, run sets MS1 and MS3, and ME1 and ME3 on the one hand, compared with MS2 and MS4, and ME2 and ME4 on the other, show F=B. We now consider the cross checks represented by collection weighting by formula F0, and by relevance weighting using F4. Unfortunately, for relevance weighting, there are some problems of control. It is possible to argue that relevance weighting may be affected by the number of relevant documents exploited for the weight calculations, so a proper comparison focussing on request membership

requires that though substitute or additional terms may come from only one or two relevant documents, weights must be calculated from all, i.e. using real R for the requests. This is somewhat artificial, but is required for test purposes. Using F1,F2,B1 and B2 respectively to define R, i.e. as sources for weighting, represents searching with a variant collection, described below. If we take subsitution with collections weights (W), run sets MSW1-5 show 1=ALL<=2, while substitution with relevance weights (R) in run sets MER1-5 gives 1<2<ALL. The cross checks thus do not give a consistent picture, but there are no large performance differences. Further, cutting across the run sets for MSW and MSR for the F/B contrast, we get F=B. For expansion with weights in run sets MEW1-5, ALL<=1=2, while for expansion with relevance weights in run sets MER1-5, 1=2=ALL. The picture here is more consistent, in that there are no real differences for 1,2 and ALL. For the expansion runs comparing F and B, again F=B overall. Thus in general (for this data), the choice of one, two or (just over) three relevant documents has no striking effects, while the choice of random or best matching relevant documents has no effect.

Now considering the comparison between substitution and the original requests, and between the expanded and original requests, i.e. between S and T and between E and T, the simple terms matching runsets MT1 and MS1-5 give S=T, while MT1 and ME1-5 give E=T. With weighting, MW1 and MSW1-5 give S<=T, while runsets MW1 and MEW1-5 give E<=T. With relevance weighting, runset MR2 with runsets MSR1-5 show S<T in general, while MR2 compared with MER1-5 shows T<E except that T has higher recall. Overall, the picture produced by taking the three alternatives T, S and E together is that they perform much the same for simple term matching, very similarly for collection frequency weighting, but that with relevance weighting S<T<E, except that S using ALL is superior to T, and that T generally has a higher recall ceiling.

Testing with the non relevance-based methods of enlarging requests was very limited. The main objective of the work was to link the project with experiments being done by Harper and van Rijsbergen. In Harper's experiments the full dependency model has been approximated by combining an MST collection derived from the collection with relevance weighting of the enlarged requests. As he kindly made available the enlarged requests for the C1400I collection, i.e. for the Even subset, it was possible to do some tests comparing this approach to expansion both with no expansion and with the alternative cruder methods using relevant documents just discussed. Further, to link the work with earlier studies of classification reported in Sparck Jones and Bates using rather cruder and more economic approaches to classification, the MST enlarged requests were compared with requests enlarged via STARS: these are classes consisting of a term and that most similar (or those equally similar) to it, obtained for the non-frequent term vocabulary for a collection using the Jaccard similarity coefficient (such classes were called Stars2 in Sparck Jones and Bates).

The initial comparisons in this group were therefore between the two methods of classification. For class enlarged term matching using run MC1 and MC2 we have MST=STARS, for classes with collection frequencies in runs MCW1 and MCW2 however we have MST<STARS, while for relevance weights in runs MCR1 and MCR4 we have MST=STARS. The comparison between class enlarged term matching MC1 and MC2 and the original request term matching of run T1 shows C<T, while that between

classes C and terms T with collection weights in runs MCW1 and MCW2, and MW1, shows C<=T. The tests with relevance weighting were limited to the use of ALL relevant documents, and compare runs MR2 and MCR1 and MCR4. They show class enlargement the same as terms, C=T. Finally, comparing class enlargement C with relevance expansion E on term matching (runs MC1-2 and ME5), weighting (MCW1-2) and MEW5), and relevance weighting (MCR1-4 and MER5) shows C<E, except at highest recall. It is possible that more experiments should have been done with the classifications, but the results did not seem worth the effort, and it was in any case felt that there was a good deal to be said for awaiting the results of Harper's own more extensive tests.

Alternative performance representations for this group of tests are few. They are available for the most interesting option comparison, namely between requests expanded with ALL, requests enlarged via the MST classification, and the original requests, all with relevance weighting by F4. Runs SrcER5, SrcCR4 and SrcR2 interestingly confirm the superiority of the simple relevant document expansion, since T<C<<E.

Variant environments

For the relevance expansion the tests under this head essentially involved the one or two relevant documents exploited to replace or enlarge requests as the source of weighting information. Thus these environments are really only significant for those cross check comparisons between terms, substitution, and expansion involving relevance weighting. The overriding effect here is of the amount of relevance information available for weighting in relation to the exhaustivity of the request. This is discussed further later in the context of the relevance weighting tests. As far as comparisons between the different treatments of request membership go, in relation to the environments represented by F1, F2, B1 and B2 respectively, runs v2/MSR1-4 and v2/MER1-4 show that substitition and expansion perform the same, i.e. S=E; but that in comparison with terms in runs v2/MR2, both S and E perform less well in general than terms, i.e. S/E<T, especially in relation to recall. However with two relevant documents available, precision at low recall is superior to that for terms: i.e. for F1 and B1 S/E<<T, for F2 S/E<T on recall but S/E>T on precision, while for B2 S/E<<T on recall but S/E>T on precision.

Experiments with class enlargement were done only for the MST classification and with environment B2. By comparison with terms, in runs v2/MCR4 and v2/MR2 we have C<T, but by comparison with relevance expanded requests in run v2/MER4 have C>E on recall but C<<E on precision. In the recall/precision performance polarisation which seems to occur in the limited relevance environments, the class-enlarged requests behave rather more like terms than they do like the relevance expanded requests in that they maintain recall at some cost in precision.

Unfortunately alternative performance representations are not available for most of these searches. The exception (due to the interest in comparisons with Harper's own results) is that a comparison can be made between terms with relevance weights and class-enlarged requests with relevance weights, for B2. This shows C<<T (runs v2/SrcR2 and v2/SrcCR4). However, as will be discussed further in Section C, the recall cutoff method is especially unsatisfactory for searches where the 'true' recall ceiling is low.

### 4.1.3 Conclusion on request membership

Overall, the comparisons just presented suggest that where relevance weighting is not used, there are no advantages in substitution or enlargement methods, whether these are relevant document-based or classificatory. However when relevance weighting is used and a fair amount of relevance information can be exploited for weighting, relevance expansion appears superior to substitution, and may be positively advantageous compared with unexpanded requests. Thus in the regular environment expansion performed better than substitution, and was better than terms where weights were calculated from several relevant documents, whether or not these were all used to expand request membership. In situations where there is little relevance information, and a requirement for high precision, both substitution and expansion techniques may have some small merit, with the latter perhaps likely to have more.

Compared with the crude relevant document strategies, there appears to be no special profit in the classificatory approaches.

Some of these points are illustrated by Graphs 1,2 and 3. Graph 1 compares the original requests, the requests expanded using ALL and the requests expanded using F1, all with F4 weights derived from all the relevant documents in the weight generation set. Graph 2 compares expansion with ALL and MST class enlargement, both with F4p. Graph 3 makes the same comparison as Graph 2, but using the alternative recall cutoff method of performance representation.

Unfortunately, the tests done by the project under the request membership heading were very limited, and more research especially on relevant document-based expansion in combination with relevance weighting would seem to be called for.

## 4.2 Request weighting

The experiments relating to this indexing factor constituted the main work of the project. the tests were divided into two groups, like those for request membership, namely those in the regular collection environments and those with the variant collections. But as the previous chapter results suggest, the variant tests were much more critical for request wighting, and specifically relevance weighting, than for request membership, since weighting is apparently more affected by the amount of relevance information available. Moreover the fact that weighting appeared to contribute far more to performance improvement meant that an investigation of the conditions for effective weight generation were especially important.

The first part of this chapter therefore describes the regular environment tests, which were concerned on the one hand with the behaviour of formula F4, previously established as effective, for different collections, and on the other with comparisons between F4 and other weighting formulae. Collection frequency weighting, regarded as a limiting case of relevance weighting, is described at the end of this chapter. The second part deals with the variant environment tests. Unfortunately, as noted earlier, though the ideal for cross checking on weighting is by changes in membership, the limited membership tests made systematic cross checking exploiting membership variables impossible, and

the relevance weighting experiments were really checked by the use of very different collections. The use of variant collections could perhaps be regarded as a further check of this sort, but is not so treated here.

### 4.2.1 Weighting formulae

Top priority in the experiments was given to evaluating the performance of F4, found satisfactory by the previous project, across the set of test collections available. Prediction was systematically from the Even weight generation document sets to the Odd weight application sets.

This strategy is a wholly appropriate way of simulating the use of weights in an SDI environment, as a comparison with the genuine SDI situation in Barker, Veal and Wyatt's 1972 tests suggests. It is also, as indicated earlier, a reasonable way of simulating an on-line searching environment. Thus it is obviously a model of an on-line situation where simple searches aiding request formulation are done on a small data base before submission to a large one. It is also a reasonable simulation of iterative searching on a single data base as a whole as long as there are a fair number of documents relevant to a request to be found in the data base: in this situation we can fairly assume that there are relevant documents to be found which are like those already inspected, say in sample output from a simple term search. The more specific value of relevance weighting as a device for extracting or promoting relevant documents not matching well in simple term searches is considered in relation to the variant environment tests.

One reason for adopting the Even/Odd strategy is that it aids experimental control in the study of iterative searching. For the strictly controlled tests relating to the amount of relevance information available discussed under the heading of variant environments, particular numbers of relevant documents may be required for weight generation. To retrieve these in a first search using simple term matching would ordinarily imply inspection of different numbers of documents per request, leaving collections of different sizes for different requests in the weighted searching. This may be unimportant for really large collections, but is objectionable for small ones since performance averaging over the different sizes of weight application set presents problems. On the other hand, if some specific number of documents is inspected for all requests in the first search, different numbers of relevant documents per request will be obtained, so weight generation is on different bases for different requests. This again may not matter much with large collections, but may present more problems with small ones. The Even/Odd strategy avoids these methodological problems and has the advantage that a whole range of test comparisons can be done in a systematic way. A second, non-trivial reason for adopting this strategy is that it is practically much more convenient and economical.

### 4.2.2 Prediction

Given that the objective of relevance weighting is to improve future searches, i.e. that we are dealing with predictions about the value of request terms, some allowance must be made for uncertainty. In Robertson and Sparck Jones 1976 a simple approach to this estimation problem was adopted, which took the form of adding 0.5 to the central contingency table elements (and 1 or 2 as appropriate to marginal

elements). Thus F4 as given above becomes

$$w = \log(((r+0.5)/(R-r+0.5))/((n-r+0.5)/(N-n-R+r+0.5))).$$

One consequence of this approach is that request terms with r = 0 are not lost, and an alternative interpretation of the approach is that it reflects the status, i.e. importance, of the original request terms.

The form of F4 given earlier is appropriate only where perfect information about the relevance distribution of request terms is available. This situation does not normally occur, but is found for test collections. So, as originally suggested in Sparck Jones 1975, and developed in Robertson and Sparck Jones 1976, for such collections F4 can be used to provide a performance yardstick. That is, if weights calculated using F4 are applied in a retrospective search of the collection from which they are generated, optimal relevance weighting performance is obtained. Thus if such weights are both calculated from and applied to the Odd subsets of the test collections, the results will show what relevance weighting performance can be obtained for these subsets and so will provide a standard for evaluating the effectiveness of prediction from the Even set. A comparison between the predictive and retrospective performance of weighted searches on the Odd set will show how good the information used for the former is.

However if predictive performance is poor, it could be that the formal expression of estimation represented by adding 0.5 etc. to the contingency table elements is unsatisfactory. The estimation problem is an extremely difficult one, and is clearly seen when very little relevance information is available, as will appear in the discussion of the variant collection tests. Unfortunately, the simple predictive/retrospective comparison just considered does not distinguish the two constituents of predictive performance very well: i.e. if predictive weighting performs less well than retrospective, it may either be because the information available for generation is poor, or because estimation from it is being done wrongly (disregarding the possibility that the weight application set is just very unlike the weight generation set, about which nothing can be done). However the possibility that estimation is being done wrongly can be investigated if the predictive version of the formula is applied retrospectively. For if the two versions of the formula give different results when applied retrospectively, having been calculated from the same relevance information, this may indicate whether the estimation technique is adequate. The comparison will unfortunately not necessarily show that the estimation technique is adequate: this is because, as Robertson has pointed out, the strictly retrospective case exploits the specific quirks of the particular body of data involved, while the formal expression of estimation is meant for general application, and so may do less well in any individual case. But we can expect to learn something about the estimation method from the comparative behaviour of the two versions of the formula over several collections. Given such relevance weighting formulae as F4, therefore, we may on the one hand do straightforward predictive experiments, primarily intended to throw light on the adequacy of the information being used to calculate the weights, and on the other retrospective searches designed to throw light not so much on the performance of relevance weighting as on the behaviour of the formulae as formal objects.

It should be noted that in applying the relevance weighting formulae in the pure yardstick mode the distributional properties of

terms may mean that particular terms are either absolutely helpful or absolutely unhelpful, either when present in or absent from a document. For example if a term is known to occur only in relevant documents, this means that any documents containing it should be promoted to the top of the search output ordering. The various possibilities for such helpful and unhelpful terms were detailed in Robertson and Sparck Jones 1976, and are reflected in the program implementation by the assignment of arbitrarily large positive or negative weights which override all others. The list of cases and consequent treatment of terms are given in Figure B1. Unfortunately, checking for such cases in yardstick runs makes these runs expensive. In the descriptions of the tests, the different applications of a formula. say F4, are referred to as F4p for the predictive application, as from Even to Odd, F4r for the retrospective yardstick application. as from Odd to Odd, and as F4pr for the retrospective application of the predictive formula, also as from Odd to Odd.

### 4.2.3 Specific formula tests

#### 1) Formula F4

The first set of experiments to be considered is the simple predictive experiments from Even to Odd sets, using formula F4. We consider first, as in the previous chapter, the main document value performance representation. The search results are given in run set MR2, for all the test collections. These show quite unequivocally that, using weights derived from all the relevant documents in the Even set, performance for the relevance weighted request terms is strikingly superior to that for the original unweighted terms. The results are very striking, the difference between the relevance weights of F4p and terms, T, being at least that represented by F4p>>T, and often that represented by F4p>>>T. The variation in degree of superiority does not appear to be associated with any specific feature of the collections, for example request or document exhaustivity, or number of relevant documents, and further collections would be required to provide more information about its possible cause. However the important point is that the improvements in performance even in the least good cases are large, and that over the set of collections, the gains made with predictive relevance weights based on F4 are some of the most conspicuous obtained in recent retrieval research, and are obtained irrespective of the very different characters of the collections involved.

We may note further that though cross checks through request membership alterations are not generally available, the improvement in performance over terms given by F4p is maintained (except at highest recall) with the relevance substituted and expanded requests for the C1400Io collections, as runs MSR5 and MER5 show.

The yardstick runs with F4r shown in run set MR6 moreover show that relevance weighting could in principle achieve even further improvements in performance compared with term searching. Thus predictive weighting from Even to Odd is substantially inferior to the yardstick, since, comparing MR2 and MR6 we have either F4p<<F4r or even F4p<<<F4r. That some of this may be attributable to the estimation technique embodied in F4p is suggested by the comparison between F4r and F4pr. The results for the latter, shown in run set MR4, are generally superior to the genuinely predictive weighting of F4p, i.e. F4p<F4pr or F4p<<F4pr, but equally the comparison between the yardstick and the

retrospective prediction given by run sets MR6 and MR4 respectively shows a general superiority for the yardstick. The results are somewhat variable, with F4pr<=F4r for the N11500Ao collection and even F4pr=F4r for the N11500To collection, but otherwise we have either F4pr<<F4r or F4pr<<<F4r. It is not quite clear why the N11500 collections show less difference: their obvious common property, short requests, is shared with the C1400I and U27000T collections.

Graphs 4-9 illustrate the comparison between F4p, F4pr and F4r for all the collections except U27000To; they quite clearly show the potential of relevance weighting, at least according to this method of performance representation.

2) Other formulae

a) Formula F1

Both Robertson and Sparck Jones 1976 and the 1977 report included comparisons between F1 and F4. As indicated earlier, the theoretical arguments were for output ordering by both presence and absence and for distinguishing term distributions in relevant and non-relevant documents; and in so far as these apply to actual data, the theoretical preference for F4, particularly in contrast to F1, should be supported by superior performance. The tests included in the 1977 report showed F4 performing better than F1 for the few collections used, namely the C1400I, U27000T1 and U27000Pb1 collections. However since the comparison is of some importance, further tests were carried out for other UKCIS collections, and also the E2500P collection. But as these results were consistent with the earlier ones, additional experiments with the E2500T and NPL collections were thought unnecessary.

The predictive search results are given in run set MR1. Performance when compared with that for terms in run set MT1 shows F1p consistently performing better than simple term searching, with F1p>T or F1p>>T. Further, retrospective performance for F1, given in run set MR5, shows F1p<<F1r or even F1p<<<F1r; but the comparisons involving the retrospective application of the predictive formula, of run set MR3, on the other hand, in general show very little difference between F1p and F1pr: i.e. effectively F1p<=F1pr, while F1pr<<F1r.

More interestingly, as these results suggest, F1 performs consistently less well than F4. The comparison between F1p and F4p (run sets MR1 and MR2) shows F1p<<F4p or F1p<<<F4p, except for the C1400Io collection, and even here F1p<F4p. Again, comparing the yardsticks in run sets MR5 and MR6 shows that F1r<<F4r or F1r<<<F4r, i.e. that in principle F4 can give much better performance, the only exception being E2500Po where F1r<F4r. Most interestingly, comparisons between the run sets MR2 and MR5 show that the predictive F4p nearly always performs the same as the yardstick F1r, i.e. F4p=F1r; indeed for the U27000Pbo collection F4p>F1r. The exception is again E2500Po, where F4p<<F1r.

F1p and F1r are compared with F4p and F4r for the relevant collections, C1400Io, U27000Pbo and E2500Po, in Graphs 10-12.

b) Formula H1

The suggestion that the method of estimating embodied in the addition of 0.5 etc. is crude was further investigated via some tests

with a more elaborate weighting formula developed by Harper (see Harper and van Rijsbergen 1978).

We can rewrite F4 as follows. We have four components, namely

(1) $\log(rN/Rn)$

(2) $\log((n-r)N/(N-R)n)$

(3) $\log((R-r)N/R(N-n))$

and

(4) $\log((N-n-R+r)N/((N-R)(N-n)))$,

which are combined to express F4 as

$w = (1) - (2) - (3) + (4)$.

(we refer to this as F4'). Harper has been investigating the following formula. We again have four components, namely

(1*) $(r/n)\log(rN/Rn)$

(2*) $((n-r)/N)\log((n-r)N/(N-R)n)$

(3*) $((R-r)/N)\log((R-r)N/R(N-n))$

and

(4*) $((N-n-R+r)/N)\log((N-n-R+r)N/((N-R)(N-n)))$,

and these are combined to give the formula

$w = (1^*) + (2^*) + (3^*) + (4^*)$,

which may be labelled H1. F4' thus has four components of the general form (logY), which are elaborated in H1 by being given the form (XlogY); and it is possible that the additional X constituents of H1, though not originally proposed as such, may be a better estimation device than the simple addition of 0.5 to F4. Some runs were therefore done to see whether Harper's revision of F4 as H1 is useful in practice.

Some of these runs are dealt with under the variant collections heading, where the comparison between F4 and H1 is particularly important. The remaining tests represent straightforward comparisons for the regular environments, for the C1400Io and E2500Po collections. The predictive search figures are given in run set MR12. The comparison with term matching in MT1 shows at best H1p>=T, while in comparison with F4, H1p<<F4p. For the C1400Io collection a cross check is possible using enlarged requests. Thus as part of a general attempt to relate the project work with Harper and van Rijsbergen's, some runs were done both with the crude STARS classification and Harper's own MST classification. The results here confirm those obtained for the original requests, that is comparisons between runs MCR1 and MCR3 for F4p and H1p with STARS show H1p<<F4p, and between MCR6 and MCR4 for MST similarly show H1p<<F4p.

These results were rather disappointing given the rationale for H1. However a more useful test of its value is for variant environments with little relevance information for weight generation; and some runs of this kind were done which are described in the paragraphs on variant collections below.

Graphs 13 and 14 illustrate the comparative performance of H1p and F4p for the C1400Io and E2500Po collections.

c) Formulae U1 and U3

An alternative approach was stimulated by the original UKCIS relevance weighting studies reported in Barker, Veal and Wyatt 1972 and in Robson and Longman 1976. In these experiments very simple relevance weighting formulae were applied, basically to obtain terms which the user might adopt to modify his SDI profile. These terms were identified by

iterated searching of the document set using a list of terms indexing relevant documents already found. In each iteration the terms from the known relevant documents were ranked by the weighting formula, and those above a threshold were selected for the next search (i.e. the weights were never used in matching). At the end of the iterative search the user modified his profile to search a new set of documents. In Robson and Longman's tests the derived list was itself taken as a profile, i.e. the terms above the threshold in the ranked list were treated as a substitute request; but the terms were still not weighted for searching. The term lists nevertheless performed surprisingly well, and this suggested that some further runs using the UKCIS formulae might be worth while.

The UKCIS workers first used a very simple formula, namely

$$w = r/n$$

which we may label U1. They subsequently studied a version designed to avoid placing terms with r = n (most often singletons) at the top of the ranked list of terms ordered by weight, namely

$$w = r^2/n$$

which we may call U2. Robson and Longman's experiments were carried out with this.

However the UKCIS project does not seem to have compared the two for performance, and arguments can be produced in favour of either. It was originally intended to compare the two formulae, but these tests were not done, and the arguments in favour of U2 when weighting is exploited automatically rather than to rank terms for the user are perhaps less strong; indeed it is possible that as in U2 small differences of term frequency are represented by large differences of value, this formula may be less effective in automatic searching than U1. The project tests were thus all done with the original formula U1. A few tests were also carried out with a modification suggested by Robertson, motivated by considerations parallel to those distinguishing F4 from F1. Thus U3 defines

$$w = r/(n-r).$$

Further, since the weights were to be applied predictively, it was thought desirable to allow for inadequate information in weight generation, so we have U1p where

$$w = (r + 0.5)/(n + 1)$$

and U3p where

$$w = (r + 0.5)/(n - r + 0.5).$$

The UKCIS staff seem not to have made such adjustments, presumably because they were primarily interested in ranking terms for the user, and because in their automatic weighted searching the profiles were so long that terms with r = 0 could be allowed to have w = 0. Again, paralleling the use of F1 and F4, we can apply U1 and U3 retrospectively to provide their own optimal performance. (It should be noted that care is needed with retrospective U3 when r=n: the weight here was taken to be r.)

The results obtained with U1 were quite unexpected, and the tests with this formula turned out to be very interesting. Experiments were therefore done with all the regular collections except U27000T.

We first compare the predictive use of U1 with simple term matching, i.e. the output given in run sets MR7 and MT1. These show U1p>>T, with an even larger improvement for U1p over terms for the

U27000Pbo collection. The yardstick runs comparing run set MR10 with MR7 show at least U1p<U1r, with U1p<<<U1r for the two Evans collections. Comparison with the retrospective application of the predictive formula were somewhat inadequately done only for the C1400Io and U27000Pbo collections, showing U1p<U1pr and U1r=U1pr, i.e. the two retrospective formulae performing the same. This comparison should properly have been made for other collections as well.

However it is the comparison between U1 and F4 which is really interesting. Considering the predictive use of the two formulae first, we compare run sets MR2 for F4p and MR7 for U1p. These show variable results, with U1p<F4p for C1400Io and the two NPL collections, U1p=F4p for the two Evans collections, and U1p>>F4p for the U27000Pbo collection. The retrospective application of the two formulae in run set MR6 for F4r and MR10 for U1r show a similar pattern of differences, i.e. U1r<<F4r for C1400Io and N11500Ao, and U1r<=F4r for N11500To, contrasting with U1r=F4r for E2500To and U1r>F4r for U27000Pbo and E2500Po. The hypothesis which these results suggest is that U1 is a good approximation to F4 for long requests. Indeed the fact that U1p does almost as well as F4p for the classification enlarged C1400Io requests (runs MCR2 and MCR5), and perhaps rather better than in the parallel comparison for unexpanded requests, is a pointer in the same direction.

The tests with F3 were extremely limited. Trial runs with predictive weighting for the C1400Io and U27000Pbo collections (run set MR8) showed U1p=U3p, and suggested that other than for very small collections (or poor relevance information) the two formulae would perform much the same. One run with retrospective weighting on C1400Io (run MR11) showed U1r=U3r. The tests with U3 were therefore not pushed further.

Graphs 15-20 compare U1p and U1r with F4p and F4r for all the collections except U27000To, showing the large performance gains obtained with U1, for this method of performance representation, with some of the collections.

Formula F0

Finally in this section we consider collection frequency weighting with formula F0.

The tests here were mainly intended to check the results obtained by the earlier project, but using more and larger collections. Thus comparing collections weights W with terms T, i.e. run sets MW1 and MT1, we typically find F0>T, though sometines, for title collections, F0>=T. A rather limited comparison between the formula using max n and the original version using N, for which results are given in run set MW1.1 show that, referring to the version using N as F0*, F0=F0*. It is possible that a more refined interpretation would suggest F0>=F0*. Overall, it appears that collection frequency weighting using F0 offers modest performance improvements.

### 4.2.4 Alternative performance representations

As the relevance weighting tests, and particularly those using F4, were central to the project work, it was regarded as essential that alternative performance representations should be obtained for the main

searches. The results obtained, as represented by the main document value method were, moreover, sufficiently interesting for it to be a matter of some importance whether the large improvements for relevance weighting were also exhibited by alternative representation techniques. These techniques were not, however, applied to the U27000T collection, as performance for this is in general so poor that the effort involved was not thought worthwhile. Further, because of the programming effort required, no full yardstick runs were done for the profile collections, which have a distinct data format. For these collections retrospective runs were only done applying the predictive formula retrospectively.

For relevance weighting using F4, the recall cutoff graphs of run sets SrcR2 and SrcT1 clearly demonstrate the superiority of F4, with F4p>>T. The yardsticks for the non-profile collections, in run set SrcR6, generally show F4p<<F4r, though for the N11500To collection we have only F4p<F4r. The predictive version of the formula applied retrospectively, for all the collections, in run set SrcR4, typically shows F4p<F4pr, though F4p<=F4pr for N11500To. Where the comparison between the two forms of retrospective weighting can be made, using run sets SrcR4 and SrcR6, we have F4r generally superior to F4pr, but varying from F4r>>F4pr to F4r>=F4pr.

The only comparison for formula F1 is reproduced from Sparck Jones and Bates; run SrcR1 for the C1400Io collection shows F1p>T and F1p<=F4p: this accords with the generally small differences for formulae for this collection. It is likely that larger differences between F1 and F4 would appear for the other collections, paralleling their behaviour with document value, so further recall cutoff runs with F1 were not thought worthwhile. But, properly, some more runs should have been done.

Rather restricted comparisons betwen predictive H1 and terms on the one hand and F4 on the other, for the C1400Io and E2500Po collections, using run sets SrcT1, SrcR12 and Src2, show at best H1p>=T for the C1400Io collection, and H1p<=F4p for C1400Io but H1p<<F4p for E2500Po.

Comparisons between predictive U1 and terms and F4 respectively for the C1400Io, U27000Pbo and E2500Po collections, in run sets SrcT1, SrcR7 and SrcR2 show T<U1p<=F4p. Retrospective performance (run set SrcR10) shows U1p<<U1r or even U1p<<<U1r. Though F4r outputs are not available for the profile collections, it may be noted that U1r>=F4pr for the three collections (run sets SrcR10 and SrcR4).

It is apparent, in other words, that in most cases the recall cutoff representation technique gives comparative results similar in relative difference, though not in absolute difference, to those given by the document value method. The rather different absolute pictures of performance given by the two methods are, of course, attributable to the methods themselves.

The formal properties of the recall cutoff method must, moreover, be responsible for the one case of material divergence in the performance pictures provided by the two methods of representation: namely the comparative performance for U1 and F4 for U27000Pbo. This collection has a number of requests of very high generality, achieving even quite low recall only at low ranks and hence depresssing average performance. If the document value method provides an unrealistically favourable view of

weighting for such collections, recall cutoff provides an unusually gloomy one. This point is discussed further in the conclusion to Section C.

To illustrate the alternative performance representations, Graphs 21-26 parallel Graphs 4-9, showing results for F4, and Graphs 27-29 parallel Graphs 15,16 and 20, comparing U1 and F4 for the C1400Io, U27000Pbo and E2500Po collections.

Unfortunately the recall cutoff runs for collection weights are very limited; those done for the C1400Io and E2500Po collections (run set SrcW1) show F0=T for the former and F0>=T for the latter. A somewhat larger performance improvement for F0 might be obtained for other collections for which the F0 performance as represented by the document value method was more superior to term performance. Such runs should have been done, but regrettably were not. It is in any case likely that the overall conclusion for collection frequency weighting using this representation method would be the same as that for the document value method, namely that collection frequency weighting at most provides small performance improvements compared with terms.

## 4.3 Variant collections

The tests with relevance variant collections were primarily intended to investigate the impact on relevance weighting effectiveness of different relevance conditions affecting weight generation and weight application.

Broadly, these conditions are of two types, qualitative and quantitative, which may merge in specific cases.

### 4.3.1 Qualitative conditions

As far as qualitative conditions are concerned, one possibility is that the relevance information available for weight generation may be quite rich, but for one reason or another the documents in the weight application set are rather different in character from those in the weight generation set, so that the request term weights obtained from the former are not very good predictors of term value for the latter. There must be some limit to this in that if the sets are so distinct that the generated weights are useless in application, this would imply that the user's view of relevance has changed radically, and with it, presumably, his query, or rather his need. An intermediate situation is nevertheless quite likely: for example we might have batches of documents from different journals in an SDI system, representing somewhat different subsets of the user's complete relevance set for the given query. In the longer term, in an SDI system, the user's view of relevance may change with time, but we can expect to adapt to this by cumulating information from the relevance assessments for successive searches, so that the request term weights are continually re-evaluated.

It is important here to distinguish query and need, since relevance is related to need rather than query. Thus even though the aim of relevance weighting is, by relying on assessments, to adapt searching to the user's implicit need rather than explicit query, if the user's view of relevance changes really substantially over a long period of SDI

operation, we can assume this reflects a change in his need which will eventually lead to a conscious modification or replacement of his profile.

In an online situation, by contrast, we assume that the user's need is constant, and it is the query which has to be manipulated to express this. Different formulations of the search can, however, be expected to select different subsets of the wanted relevance set, so the situation is similar to the short term SDI one (though much more difficult to investigate).

The project has not been carried out in an environment allowing systematic investigation of this aspect of weighting. Thus we did not have any temporally characterised or phased sets of relevance assessments. It was, however, possible to do some limited tests relating to the effectiveness of weighting for heterogeneous relevance sets. Thus the fact that the UKCIS data consists of two sets of documents taken from CAC-1 and CAC-2 respectively and representing rather different subject areas of chemistry allowed a test of prediction from one to the other. Most queries in the test data have relevant documents in both sets but these are usually unevenly divided. The predictive search result could therefore suggest whether relevance weighting could work in what may be described in a hospitable manner, though more extensive experiments would be required to demonstrate this.

The UKCIS collection was also large enough to exploit for a crude simulation of relevance weight consolidation over time. Thus the nested FRACTION subsets, Threequarters, Half, Quarter, etc., could be used in reverse, as it were, to see how prediction becomes more reliable as a better spread of relevance information is obtained. However it must be said that as the subsets were established by pseudo-random selection, their use in this way is primarily to test the effects of the sheer quantity of relevance information available on weighting performance, i.e. in relation to the quantitative aspect of weighting. These tests can be treated as ones on the quality of relevance information only on the assumption that there is a good deal of content variation in the documents relevant to a request (particularly if different grades are not distinguished), so that increasing the quantity of relevance information is not merely a matter of reinforcement of existing information, but one of improving the quality, i.e. relevance coverage, of this information. But as the number of relevant documents involved in the FRACTION sets increases quite rapidly, the tests with these cannot be taken very seriously as coverage tests, and are more properly considered as quantitative. We can at best treat the experiments with the smallest subsets as relevant to the qualitative aspects of weighting.

Another qualitative aspect of weighting is the degree of relevance of the documents involved: the discussion so far has assumed either that the relevant documents are equally relevant, or alternatively that their degree of relevance is immaterial. However it might be supposed that highly relevant documents might be particularly useful for prediction, so in iterative searching highly relevant documents might be selected from the relevant documents found in any one cycle for use in generating the next cycle's weights. Alternatively, documents with high matching scores in one matching cycle may be selected for weight generation in the next: highly matching documents are not necessarily highly relevant documents, but may be expected or assumed to be. The

analogy with the document set case becomes clearer if we look at individual document quality the other way round: i.e. if we assume that the user is most interested in highly relevant documents, can we retrieve these effectively even if the relevant documents in the weight generation set are not, or are not all, themselves highly relevant?

The project did not explicitly test for the value of highly relevant documents for weight generation, or for the effectiveness of relevance weighting as a means of selecting just highly relevant documents; but it did investigate the use of highly ranking documents for weight generation, in a simulation of on-line searching. The failure to study the explicit use of highly relevant documents, or the ability of relevance weighting to select them, given that appropriate data for some collections is available, must be deemed a gap in the project. But the tests with highly matching documents can be claimed as more important, since they directly reflect the basic conditions of online searching.

### 4.3.2 Quantitative conditions

The quantitative side of weighting was investigated firstly through the comparisons between systematically enlarged FRACTION weight generation sets for the UKCIS collections, and secondly through experiments deliberately restricting the numbers of relevant documents available for weight generation.

In the first group of quantitative tests the weights were generated using full information for the document set exploited, i.e. using all the relevant documents known for the set. (The fact that the assessments for some of the collections were not exhaustive was disregarded.) The tests were thus unrealistic in that it cannot be assumed that in practice all or even most of the relevant documents in the weight generation set will be readily found. The objective of the tests was a rather more formal one, namely to establish how effective relevance weighting can be given little but accurate information for weight generation.. The particular point of interest was to observe changes in the performance of the weights with the increase (or reduction) in the amount of relevance information exploited to obtain the weights.

As indicated in Section A, in the smallest FRACTION set, Sixteenth, the average number of relevant documents per request, for the U27000T collection with 182 requests, is 3.7, compared with 44.2 for the largest set, Threequarters. The actual number of relevant documents per request of course varies, and one object of the second set of tests was to control the numner exploited for weight generation more exactly, by using the same specified number of relevant documents for each request. Thus the tests compared weight generation from just 1, just 2, and (for the larger collections) just 3, relevant documents. At the same time, the first set of tests could be regarded as a simple simulation of the more realistic situation where there is no guarantee that all (or even most) of the relevant documents will be readily found, in that it may be supposed that at least as many relevant documents as those exploited for the smaller FRACTION subsets would in fact be found without much difficulty in an ordinary environment. The second set of tests on the other hand investigated the specific problem posed by real environments, namely that those relevant documents identified are only a sample of the ones existing, and so may not be a particularly reliable basis for weight

computation.

These tests indeed show how the qualitative and quantitative aspects of relevance weighting are intimately linked, since where there are few relevant documents, they may well be idiosyncratic: this may be true even if the known relevant documents are all or most of them to be found in the document set concerned, but it is much more likely if a random but small sample is used, or one selected by some criterion like matching rank, which could introduce bias. It is extremely difficult to determine the relative contribution of the qualitative and quantitative factors in this type of test. However it seems most appropriate to regard experiments with a few selected relevant documents as primarily quantitative ones since it is the quantity of relevance information available which may make it qualitatively unrepresentative. For the project tests in particular, though best matching relevant documents were used, where a correlation between document rank and quality might be presumed, there was no independently established qualitative characterisation of the relevant documents referring either to their general subject orientation or specific content; the selected documents could thus be arbitrarily unrepresentative. The sampling tests involved two different methods of obtaining samples. In one case, relevant documents for weight generation were obtained simply by taking the first 1, first 2, or first 3 relevant documents known as weight sources, and deeming this, perhaps too sloppily in the case of two and even more three, to be equivalent to a random selection. These choices, collectively labelled FIRST, can be taken as a simulation of the situation where the user comes to the search with some known, but not necessarily especially good, relevant document(s), or does a simple Boolean search and takes the first relevant document(s) encountered on scanning its unranked output. In the second case relevant documents were obtained by taking the highest ranking 1, 2, or 3 relevant documents retrieved by a simple term coordination search of the document set, providing BEST generation sets. As will be discussed more fully below, both FIRST and BEST were selected from the Even document sets to apply to the Odd. (There are of course other sampling bases than those considered here, representing other types of user input, and other approaches to matching.)

It is perhaps useful to summarise the weight generation factors just discussed as follows. We have two qualitative contrasts relating to document sets and individual documents respectively: in the first case between narrow and wide relevance coverage, i.e. between a specialised and neutral document set characterisation; and in the second between high and (high plus) partial relevance grades, i.e. between a specialised and neutral individual document characterisation. Quantitatively, and similarly dividing a continuum, we have a contrast between few relevant documents and many. Thus if we consider first the more formal case where the test data allows us to talk about using all, or accurate, relevance information in the weight generation set, this gives us the options in the two-part table below: in this the particular data set names used by the project are given as option labels, and the options tested are underlined.

If we take the FRACTION sets and the entire collection, ALL, to compare few and many, but otherwise unspecialised, relevant documents, we can contrast these on the one hand with the subject specialisation represented for the test data by the UKCIS CAC sets, and on the other

| | | QUALITATIVE | | |
|---|---|---|---|---|
| | | neutral w.r.t. document set & individual document | specialised w.r.t. document set | specialised w.r.t. individual document |
| QUANTITATIVE | few | FRACTIONS | CAC + FRACTIONS | HIGH + FRACTIONS |
| | many | ALL | CAC | HIGH |

with relevant document grading represented by a selective use (hypothesised as for both weight generation and user reading) of highly relevant documents only. (Following the general style of the 1977 report, these can be labelled HIGH.)

Other versions of the table then follow from the relevance sampling alternatives, i.e. from the use in the experimental context of only some of the complete set of relevant documents known for the weight generation document set. (In real life the set of relevant documents identified in some particular way, say by a search, has to be treated formally as if it did consist of all the relevant documents in the collection searched, though we accept that in fact there are likely to be others.) Thus one sampling mode is based on taking (deemed) random relevant documents, the other on taking the best matching relevant documents retrieved by simple term matching. Of course if we are dealing with a large collection we may consider sampling in relation both to the use of few and many relevant documents for weight generation, but for test purposes, with the size of collection available, the sampling options apply only to the situation where few relevant documents are used for weight generation. Thus again using the project versions for illustrative purposes, i.e. with the sampling alternatives FIRST and BEST respectively, we have:

| | | QUALITATIVE | | |
|---|---|---|---|---|
| SAMPLING | | neutral w.r.t. document set & individual document | specialised w.r.t. document set | specialised w.r.t. individual document |
| QUANTITATIVE | few | FIRST<br>BEST | CAC + FIRST<br>CAC + BEST | HIGH + FIRST<br>HIGH + BEST |

As the tables indicate, the tests of the various options possible were only begun by the project, and much more work needs to be done in this area. The experiments that were done, and especially those with FRACTIONS and FIRST and BEST, were nevertheless the obvious ones to start with, being those necessary to check relevance weighting in more obvious environments: there was little point in proceeding further if these had been unsuccessful. These were experiments primarily concerned with the quantitative aspect of relevance weighting but, as suggested above, they may also be regarded as probably having some bearing on qualitative questions.

There were thus two groups of tests corresponding to the use of full and sampled relevance information and chiefly emphasising the quantitative aspects of weighting..

As noted above, the motivation in the second case in particular was to study relevance weighting in a simulated online environment. This raises, as discussed earlier, the methodological problems presented by genuine iterative searching, and to the adoption of the Even/Odd strategy. That is, to the use of two distinct subcollections for weight generation and searching, rather than to cycling over the same set. The latter, alternative, approach has been adopted by Harper and in recent experiments conducted by Robertson, van Rijsbergen and Porter, and an attempt will be made to relate the project tests to these other experiments in Section C. The online simulation done by the project is thus a double simulation, first in not being online with real users at all, and second in treating the Even/Odd strategy as a substitute for genuine iteration.

It has to be admitted that there are methodological arguments for both approaches in relation to controlled experimentation, and some tentative iterative tests were therefore done with the C1400I collection. The results obtained were not very successful, but it was hypothesised that this was probably due to excesive 'creaming', i.e. to the fact that with only a few relevant documents to be found, and the best matching third or half used for weight generation, it may be too difficult to ugrade the remainder. It is true that, against this, relevance weighting might be looked to as a way of improving the output rank of relevant documents poorly matching on unweighted terms, but if the total number of documents involved is small, it may be difficult to achieve much, or even to perceive average changes in performance. Comparisons with iterative experiments done by Robertson, van Rijsbergen and Porter on the larger NPL data are therefore of interest, and are considered in Section C.

### 4.3.3 Specific condition tests

#### a) Document set tests: qualitative

The subject-oriented tests done under this head using the CAC subsets of the UKCIS data were relatively straightforward. The First subset, representing CAC-1 documents, was used to generate weights which were applied to the Last subset, representing CAC-2. (The names First and Last refer to the document set numbering, as described in Section A.) The runs were done for both the title and profile versions of the UKCIS data, and included searches with both formula F4 and formula F1. The various search results for relevance weights can be compared on the one hand with simple term searching, and among themselves, and on the other with performance using the Even/Odd subsets.

Considering F4 first, using runs sets MT1, MR2, MR4 and MR6, we have F4p>>T, F4p<<<F4r, F4p<<F4pr and F4r>F4pr, i.e. a pattern of relationships for the two collections, U27000T1 and U27000Pb1, the same as for the regular collections U27000To and U27000Pbo. The runs using F1 (run sets MR1, MR3 and MR5) show F1p>T, F1p<<F1r, F1p=F1pr, and F1r>>F1pr, while at least F1p<<F4p, F1r<<F4r, and, finally F4p=F1r: this pattern is again like that observed for the regular collections.

Unfortunately the Last/Odd comparison can only be made for the U27000Pb collection since the Last and Odd versions of the U27000T collection have different numbers of requests; however for the reasons already mentioned, the profile results are more important. These show that as far as F4r and F4pr are concerned, Last=Odd, i.e. as one would expect, retrospective performance is the same, and this holds for F1r and F1pr too. However for F4p Last<Odd, i.e. the prediction from First to Last is not as good as that from Even to Odd. The important point, however, is that the weighting still gives a large improvement over terms, i.e. F4p>>T, as noted above. For F1p we only have Last<=Odd, reflecting the generally poor performance for F1 (and again here, less interestingly, an improvement over terms).

There are no alternative performance representations for these runs.

b) Document set tests: quantitative

As described in Section A, a quarter of the collection, Search Quarter, was used for weight application, while subsets of the remainder representing respectively a Sixteenth, Eighth, Quarter, Half and Threequarters of the full collection were used to generate weights. Both the U27000T and U27000Pb collections were used for these FRACTION tests (except that, due to processing inertia, there was no Threequarters subset for U27000Pb).

The runs involved only relevance weights using F4, and were designed simply to study the effects of gradually increasing (or gradually diminishing) the amount of information available for weight generation. As indicated in Figure A7, the number of relevant documents and documents respectively ranged from 3.7 and 1711 for Sixteenth to 37.4 and 20521 for Threequarters (the latter is of course larger than Even, which is comparable with Half).

The comparisons are on the one hand with respect to terms, and on the other between the various subsets. Predictive performance for the subsets was also evaluated by comparison with retrospective searching on the application set.

The term runs on Search Quarter for the two collections are given in run set MT1, and the retrospective relevance weighting results for F4r and F4pr respectively in run sets MR6 and MR4. The predictive runs are in v1/MR2.

The results for U27000T, even with the smallest set, Sixteenth, show a substantial improvement in performance for the relevance weights compared with terms, i.e. Sixteenth F4p>>T, except that there is s slight loss of recall. Further, each successive enlargement gives either no, or only a small, improvement for F4p compared with its predecessor, though

it it is possible to say that Sixteenth<Quarter and Sixteenth<<Half. Moreover Half and Threequarters perform the same, and only slightlv less well that a retrospective search using the predictive formula on Search Quarter: i.e. Threequarters F4p<=F4pr.

The results for the U27000Pb collection are very similar, though absolute performance is of course enormously better. In this case relevance weighting from Sixteenth again performs much better, indeed very much better, than terms, i.e. Sixteenth F4p>>>T, apart from a slight loss of recall. The successive enlargements then quickly reach the recall level of terms and themselves perform the same. However they perform somewhat less well than the retrospective search with the predictive formula, e.g. Half F4p<F4pr, and hence Half F4p<<<F4r, as F4pr<<F4r for this collection.

For this data a trial run was done with formula F1, using Eighth. This showed F1p>T (and so F1p=W1); F1p<F1pr and F1pr<<F1r. The small weight generation set is not effective here, but this result is in accord with others for F1: a very crude, since not strictly legitimate, comparison with the Even/Odd search result shows that predictive performance for the latter is not greatly better than that for Eighth. (The comparison is not quite proper since the weight application sets are of different sizes.)

The conclusion which naturally follows from these tests is that F4 can be a very powerful device even when relatively little, but accurate, information is available for weight calculation.

It is a pity that these experiments involved only UKCIS data; further runs would be desirable, and would also justify the cost of obtainining alternative representations of performance, which were not obtained for the UKCIS collections.

c) Document set tests: sampling

The results obtained in the tests just described were very encouraging in suggesting that relevance weighting, especially using formula F4, may improve performance even when not merely little, but perhaps unrepresentative, information is available about relevant documents in the weight generation set.

As described earlier, the two approaches to sampling were the use of BEST, the best matching relevant documents derived from simple coordinate term searches of the Even document set, and of FIRST, the first documents in the given relevance lists for the Even set, assumed equivalent to random sampling. As noted, the use of the first documents was intended primarily to act as a control on the test with the best matching, though, as mentioned earlier, it could also be regarded as a simulation of the situation where the user approaches the system with a known relevant document, so that an initial term search may be unnecessary.

As well as being experiments bearing on weight generation, these tests were linked with those on request membership described in the previous chapter, where requests were replaced by, or expanded with, terms taken from the best matching or first relevant documents.

Formula F4

Like the tests on request membership, these sampling experiments cover a range of experiments. We have comparisons between weighted and unweighted terms, between the use of 1,2, or 3 relevant documents, between the use of these samples and that of all the relevant documents in the weight generation set, and between the use of first and best matching documents.

It will be evident that the range of options to be tested is very large, and it has proved impracticable to carry through a really comprehensive range of experiments. Indeed the results obtained in some cases, as will be detailed later, have made it clear that further work on the weighting formula, and specifically on the way estimation is expressed, is required before further testing in some environments is appropriate. The experiments involved all the collections except U27000T, but were predominantly devoted to studies of BEST rather than FIRST since, as noted above, the online simulation represented by the use of BEST was thought to be particularly important. However as it appears that as the size of the sample is increased any differences in performance between FIRST and BEST disappear, there is no real need for extensive tests with both. Further, initial tests suggested that using only 1 relevant document was not very effective, so effort was concentrated on using 2, and also 3 for the larger collections: but essentially BEST2 and FIRST2 were thought of as 'typical' small samples. Tests with FIRST1, i.e. F1, and with BEST1, i.e. B1, were done for the C1400I and U27000Pb collections only. Tests with B2 were done with all the collections, but with F2 only for the C1400I, U27000Pb and E2500P collections. As mentioned earlier, it was thought futile to investigate 3 relevant documents for the C1400I collection; however tests with B3 were carried out for all the other collections, though with F3 only with the U27000Pb collection.

It turns out to be virtually impossible to give the results for the different collections a general characterisation. They divide into two groups for collections with, respectively, short and long requests. The reasons for this difference are analysed in Section C. In the meantime it can be simply stated that predictive relevance weighting using F4 in the standard form applied in the project experiments, with small relevance samples and long requests, i.e. for the U27000Pb and two E2500 collections, leads to a substantial or very substantial lowering of the recall ceiling reached with unweighted terms. The description of the search results which follows therefore treats the short request and long request collections separately, on their own terms.

Short requests

Taking the short request collections first, we initially compare predictive relevance weighting with F4 based on 1, 2 or 3 relevant documents, for FIRST and BEST respectively. The smallest sample, 1, or 2, is then compared with terms, T, and the largest, 2, or 3, with using all the relevant documents in the Even set, labelled ALL. The term run set is MT1, the run set for ALL is MR2, and the run set for the various FIRST and BEST options is v2/MR2.

For the reasons mentioned, the comparison between F1 and F2 is limited to the C1400Io collection, where we find F1<=F2; we also find F1>>T and F2=ALL. Thus for FIRST, for the rather inadequate one

collection comparison, the general picture is of litle performance improvement with the increase in the number of relevant documents.

Turning now to BEST, for C1400Io we have B1=B2, with B1>>T and B2<=ALL, while for N11500Ao and N11500To we have B2=B3, at least B2>T, and B3<ALL. The picture for BEST is thus similar to that for FIRST, with variations in the size of relevance sample not affecting performance very much, and with even a small sample providing a considerable performance improvement compared with terms, and approaching that given by ALL.

When performance for FIRST and BEST is compared, for C1400Io, we find FIRST=BEST for F1 and B1, and F2 and B2, respectively.

Long requests

The situation for the U2700Pb and two Evans collections is very different. As indicated above, the conspicuous feature of the relevance weighting searches based on small samples is the lowering of the recall ceiling compared with that reached in simple term matching: roughly speaking the ceiling can be described as low if it is only 2/3 as high as that reached by unweighted terms, and very low if it is only 1/3 as high. Thus while comparison of the usual kind between performance for different samples is legitimate, because their recall ceilings are not too different, it is hardly possible, if recall is regarded as important, to make comparisons between these and either terms on the one hand or full relevance weighting on the other. When the latter is attempted, the only remark possible is that sample performance is inferior. However if a loss of recall is accepted, comparisons can be made for precision below the highest recall level common to the options being compared. This is what is done here. The interesting question raised by these results, their causes, and possible reactions to them, were first treated in Sparck Jones 1979a, and are discussed in detail in Section C.

In describing these results the convention will be adopted of enclosing comparisons subject to recall ceiling restrictions in parentheses: thus (F1>T), for example, means that F1 performance, i.e. effectively precision performance, is better than that for terms, subject to a low or very low ceiling, as indicated in the text. (Parentheses are of course not needed if all the items compared have an equally low ceiling.)

Then making these essentially precision-oriented comparisons we first consider the alternatives of 1,2, or 3 FIRST relevant documents for weight generation. These can be compared for U27000Pbo, and run set v2/MR2 shows that F1=F2<=F3, at very low recall. F1 is itself superior, but only below its ceiling, to terms, i.e. (F1>>T), while F3 is inferior in precision as well as recall to ALL, i.e. (F3<<ALL). The comparison between F2 and terms for E2500To shows (F2>>T), and between F2 and ALL shows (F2<<ALL).

As far as BEST is concerned, for U27000Pbo we have B1<B2<=B3, with B1 having a very low ceiling and B2 and B3 a low ceiling; B1 is better than terms below its ceiling only, i.e. (B1>T), while B3, with a low ceiling, is inferior to ALL in precision as well, i.e. (B3<<ALL). For E2500To and E2500Po we have a similar situation, with B2<=B3, with B2 with a very low ceiling and B3 with a low one; again (B2>T) while (B3<<ALL).

For these collections comparing FIRST and BEST over the range of samples for U27000Pbo we have F1>=B1, with B1 a very low ceiling and F1 a low ceiling, F2=B2, and F3=B3. For F2 and B2 for the E2500To collection we have F2=B2.

The general picture produced by these results in other words is that for precision, relevance weighting improves on term performance even for very small samples, though it is not as good as relevance weighting with full information; but that for recall, with long requests, there is a substantial loss. Interestingly, even with only 1 relevant document in the sample performance gains are achieved, while 2 are better than 1, but 3 only a little better than 2: though this was not investigated, there would presumably be a gradual improvement as the sample is enlarged for those cases where 3 is less good than ALL; however we may expect, on the basis of the other results, that performance as good as that for ALL will be achieved before the full sample size is reached. Comparing FIRST and BEST, it appears that the former have a slight edge when only very few relevant documents are available, but that the two converge rapidly in performance with increaasing sample size.

Formulae U1 and H1

Some more limited comparisons with weighting formula U1p were done, to see whether this formula is affected in the same way as F4p. The results are given in run set v2/MR7, for comparison with terms (run set MT1) and with variant collection F4 weighting (run set v2/MR2). The output for B3 (or B2 in the case of the C1400Io collection) shows, not surprisingly, a smaller improvement over terms than when weighting is derived from all the relevant documents in the weight generation set, ranging from U1p>=T for the NPL collections to U1p>>>T for the U27000Pbo collection, with U1p>T in the other cases. However the comparison with F4p generally shows U1p<<F4p for short requests and (U1p<F4p) for long requests, i.e. lower precision for U1p, but for long requests only below the recall ceiling for the F4 weights: for these requests recall for U1p is as high as for terms. For the U2700Pbo collection even precision is as good for U1p as for F4p. The comparison for F2 and F3 for the E2500Po and U27000Pbo collections respectively shows similar results similar to those for the BEST variants. In other words, the U1 weighting is not affected by the variant collections in the way F4 is: performance is lower than with many relevant documents in the weight generation set, but recall does not fall for long requests.

A single run for H1 with the C1400I collection shows B2>>T and also, inexplicably, B2>ALL.

There is little point in providing large numbers of graphs to illustrate these variant collection results, since the Run Tables here speak for themselves. The main points only are therefore illustrated by Graphs 30-35, for all the collections except U27000To, showing comparative performance for U1p and F4p based on B3 (or B2 for C1400Io).

4.3.4 Cross checks and alternative performance representation

Cross checks on all variant collection performance were relatively limited, being confined to runs on the C1400I collection with substituted or expanded requests, i.e. with sample relevant documents being used for both request modification and weighting. The searches

were, however, very interesting since the longer requests involved lower the recall ceiling for weighting though recall performance for the original short requests was quite satisfactory. The search outputs are in run sets v2/MSR1-4 and v2/MER1-4. Substitution and expansion, S and E, behave virtually identically. It gthen appears that comparing 1,2 and ALL, recall is very low for F1 and B1, and low for F2 and B2, so considering primarily precision comparisons we have (F1>>T), F2>f1, and (ALL>F2), and the same for BEST. The comparison between FIRST and BEST shows that recall is better for both F1 and F2, with no difference in precision between FIRST and BEST.

Given the problem of the recall ceiling, the use of recall cutoff representations for these searches must be suspect. However for the record, since very abstract arguments for the legitimacy of the recall cutoff method can be produced even in this situation, it may be noted that, as run set v2/SrcR2 shows, B3 is generally very superior to terms, i.e. B3>>T (though for C1400Io B2>T), while using ALL leads to little further gain in performance, i.e. at best ALL>=B3, and sometimes ALL=B3.

## 5 Iterative searching

As described earlier, the tests on the variant collections using the Even/Odd strategy were intended as a simulation of iterative searching. Some tests were also done with genuine feedback, in the general style of Harper's 1980 experiments.

These tests were rather tentative because the retrieval programs appropriate to experimental (though not necessarily operational) feedback searching are rather costly: explicit output ranking is needed for each search cycle. The tests all followed the same pattern, with a first search using simple term matching followed by inspection of the documents retrieved above a cutoff to identify relevant ones, the use of this information for weight calculation, and one or more further iterations to obtain output and revise the weights. Various combinations of cutoff and cycle were tried: for example 25 x 2 represents a cutoff after 25 documents in each cycle, and one cycle for term matching and one subsequent one exploiting weights; 5 x 5 represents a cutoff of five documents for each cycle, and one term searching cycle followed by four weighted ones.

There are some awkward evaluation problems with these strategies: should one compare performance for the whole concatenated series of search outputs, or exclude the first cycle output on the grounds that it is common to every strategy, evaluating only over the concatenated output of the remaining cycles. If a 25 x 2 scheme is applied, for example, and two weighting formulae are being compared with one another and also with simple term matching as a baseline, this implies removing the top 25 documents in each case, as common to all the strategies, and then comparing performance for the 26th-50th ranking documents output by the term search with the 25 documents retrieved in the second cycle, but first weighted search, for the two relevance formulae. Decisions on this point can make a good deal of difference to the overall picture of performance, as most strategies, even the simplest, retrieve a fair number of relevant documents in the top ranks.

The project tests were carried out with the C1400I and also the old C1400T collections. Overall, the results seemed disappointing, though this was perhaps a hasty judgement. The results obtained can be illustrated by the comparison between term matching and relevance weighting using F4, i.e. F4p, for the 25 x 2 case, i.e. for ranks 26-50. (It should be noted that the weights are calculated using R and r values derived from the documents above the cutoff, while N and n refer to the whole collection.) For the C1400I collection, with 25 x 2, excluding those relevant documents retrieved in the top 25 ranks, term matching retrieved 20% of the remainder by rank 50, while F4p, using weights derived from the information supplied by the relevant documents in the top 25 ranks, retrieved 29% of the remaining relevant documents. Thus from one point of view F4p is doing much better than terms, since the gain in recall for F4p over terms is 45%, but from another point of view neither strategy is doing very well overall. Comparative performance for the 12 x 2 case gives terms 16% 'relative' recall and F4p 22% by rank 24, an improvement with F4p of 37.5%.

One problem for the experimenter in such tests is the influence on performance of requests not retrieving any relevant documents by the first cutoff point. In real life the user would presumably scan until he found a relevant document, or at least might do so, but allowing a variable cutoff across requests represents a lack of control in testing. One possible way of dealing with this problem of non relevant-retrieving requests is to eliminate them after the first search, so evaluation is based only on requests actually exploiting any relevance information. This procedure has been followed by Harper. Applying it in the 12 x 2 case gives recall 16% and 23% respectively for terms and F4p, slightly increasing the gain for the latter to 43%. There are, however, some obvious objections to this procedure, and it is perhaps of more interest that the original technique showed a significant improvement for relevance weighting, even though 44 requests out of 225 do not retrieve any relevant documents through term matching by rank 12, and 21 do not retrieve any by rank 24.

As mentioned, the difficulties of conducting these iterative experiments led to the adoption of the alternative variant collection technique. But it must be admitted that more tests with the genuine iterative strategy, especially with other collections with more relevant documents than the Cranfield ones, should have been carried out. In particular it is desirable that the specific effectiveness of relevance weighting as a device for promoting those relevant documents matching only poorly on unweighted terms, as opposed to those matching well, should be further investigated.

## SECTION C : DISCUSSION

In Section B the motivation for the tests carried out by the project was described, and the results of these tests were simply presented. In this section the results are first evaluated, and then compared with relevant findings by other projects. The section concludes with a brief assessment of the project results as a whole.

### 1 Analysis of the test results

Following the presentation of the experiments in Section B under the two indexing factor heading request membership and request weighting, the project tests will be correspondingly evaluated under these heads.

#### 1.1 Request membership

As noted at the end of the presentation of the test results on request membership, the experiments carried out by the project were rather limited, so that any overall conclusions reached must be tentative. Insofar as the tests suggest any conclusions these are that request expansion may be superior to substitution, presumably because the user's original request terms are retained, and that expansion with relevance weighting may be superior to unexpanded weighted requests. However it would appear that in this case reasonably full information is required for weighting, for the reasons connected with estimation discussed more fully below. An interesting point is that simply taking terms from relevant documents could be as effective as the more sophisticated approach using a classification.

The point of interest connected with request expansion is that a very high level of performance is in principle attainable. This is shown by a trial comparing MST-expanded requests and yardstick weighting by formula F4 with the yardstick for the original requests, for the C1400I collection. The two sets of figures are given in Other Table runs 01 and MR6 respectively, and clearly show that expansion gives an enormous improvement in performance: i.e. referring to the expanded version as MST+F4r we have MST+F4r>>>F4r>>>T.

Work in this area has been carried out by Harper, so the possibilities and problems of query expansion will be considered further in the discussion of other research.

#### 1.2 Request weighting

The important points here are the general conclusions to be drawn about relevance weighting, specifically by using Formula F4, about its effectiveness in environments where little relevance information is available, and about its effectiveness compared with that of the other weighting formulae tested.

The project tests show quite unequivocally that predictive relevance weighting using the theoretically recommended formula F4 can be extremely effective: this is a consistent result across all the very

different test collections used, and across different methods of performance representation. The tests also show that prediction from accurate information, especially a reasonable amount of it, can be nearly as effective as retrospective weighting using the formula in its predictive version, F4pr.

The main problem is therefore the estimation one. That this may be a general problem, though not always a serious one, is suggested by the fact that retrospective searching using the predictive version of the formula may be much less effective that the pure yardstick; but the problem is shown up more clearly by the collapse of the recall ceiling when predicting from small samples for long requests.

The explanation for the loss of recall appears to be that documents are over penalised for not possessing the great majority of the many request terms. In detail, the situation can be analysed, as in Sparck Jones 1979a, as follows.

The relevance weighting function F4 as presented in Section B orders the search output for a particular request; but, as originally noted in Robertson and Sparck Jones 1976, the actual matching values obtained for documents as the sum of w's for query terms they contain do not give a complete picture of any document's status. To obtain complete, fully significant matching values, i.e. document scores directly reflecting relevance probabilities, scaling is required. This is also required for averaging across requests by matching values rather than output rank position.

The approach to weighting underlying F4 is based, as indicated in Section B, on the idea that both the presence and absence of search terms in a document is important. That is, we want to accept documents containing good terms, and reject ones containing bad terms; and we also want to reject documents without good terms, and accept documents without bad terms. So for a request and document with several terms, the final result is the net balance for good and bad terms.

As theoretically presented in Robertson and Sparck Jones, this approach leads to term weights with explicit presence and absence components, contributing respectively to document scores according to whether each term is or is not present in the document. The presence component is

$$v = \log((r/R)/((n-r)/(N-R)))$$

and the absence conponent is

$$u = \log(((R-r)/R)/((N-n-R+r)/(N-R))).$$

The score for a document is thus the sum of the v's for all the request terms present in the document plus the sum of all the u's for the request terms absent from the document.

The presence/absence form of F4 weighting is reinterpreted, as described in Robertson and Sparck Jones, by F4 as given earlier (without the additions of 0.5) in the form of a presence only weight w which achieves the correct output ordering through the positive or negative weights it assigns to the request terms. Good terms have positive weights and bad ones negative weights, so the document score is simply the sum of these weights. The effect of scoring for absence is achieved indirectly through the non-contribution of term weights to scores, but to obtain numerically significant scores, i.e. exactly those given by the

explicit presence/absence form of the weighting scheme, the simple sum of w scores must be modified. This is achieved by calculating a scoring constant for each request which is applied to every document. This constant is the sum of u's for all the search terms. The net effect, therefore, is to modify the contribution to a document score of any matching term, and to make a contribution for any non-matching term. The relationship between the two techniques for arriving at the correct score for a document is indicated by the fact that for a single term w=v-u.

Significant scores are also required to determine cutoff in searching. In simple term matching with output ordered by coordination level, the levels have a clearcut interpretation, and a natural cutoff is provided by level 0, representing no match, and also the minimum score. (This is a logical cutoff, not necessarily related to a human one.) Simple weighting schemes like that represented by collection frequency weighting with F0 have a similar natural cutoff at 0, and in ordinary collections it is at level 0 that most of the collection documents will be found.

The natural cutoff for the relevance weighting scheme is rather different. Specifically, a document's score reflects its probability of relevance. This is not true of simple term searching. So in relevance weighting by F4 a document score of 0 directly defines that point where the probability of retrieving a relevant document is that of retrieving one at random from the whole collection. A score of 0 does not mean that request and document share no terms, and it is not the minimum possible score. However in practice we can expect that most documents will fall below it. Thus the important point about relevance weighting scores is that any class of scores, for example positive or negative ones, do not reflect request/document matching in some simple way: positive scores reflect one particular outcome for combined v's and u's, which may individually be either positive or negative, and the same applies to negative scores. In practice, the great majority of negatively scoring documents will be those not sharing ay terms with the request, but a negative score does not imply that this is the case. It will be clear that the number of documents having a score of exactly 0 can be expected to be small.

Now the previous search results for collections with long requests have a low recall ceiling, which implies a high cutoff point in the output ordering. As just indicated, the cutoff marks the point at which scores are defined as equivalent to searching at random; and as this cutoff is defined in terms of the estimated values of the various request terms, the low ceiling implies there is something wrong with the estimation technique.

It must be emphasised that estimation is being used here not to refer to the whole idea of probabilistic weighting, which is estimating relevance, but to the specific methods of estimation exploited to derive weights from one body of documents for application to another. For the predictive application of F4, the method adopted has been the addition of 0.5, and the results obtained for the variant collections, i.e. for situations where litle relevance information is available, suggest that this method is inadequate. Its deficiencies are concealed either when good information is available, and are exacerbated by long requests when little information is supplied.

More specifically they appear to be exacerbated when long requests containing terms with particular collection properties are being used. These are terms which do not occur in any of the relevant documents known, i.e. terms which have r=0; in the case of profiles in particular, which frequently contain a range of spelling variants etc., such terms may indeed have r=0 because they do not occur in the document set at all, i.e. they have r=0 because n=0. Further, with long requests there may be many such terms, so that their aggregate impact on any document score can be large. Specifically, the estimation technique used in the tests will give large positive w and small negative u for terms with r=0 and n=0, allowing future documents containing such a term to score positively, and documents not containing it to score negatively. Unfortunately, if there are many such terms the net result will be that many documents are heavily penalised with negative scores, so that they are not retrieved. Some relevant documents are clearly lost in this way, with any contribution from a positive, but not strongly positive term, overwhelmed by the combined contributions of many negative, even if not strongly negative, terms. The fact that all the test collections which have long requests have short documents must contribute to this outcome, since the effects of the negative terms are not counterbalanced by matches on several positive terms. Thus it is only those documents which have a very positive request term or terms which succeed in being retrieved with an overall positive score. These are of course highly likely to be relevant, accounting for the fact that while the recall ceiling is lowered, precision is improved.

Overall, therefore, the position seems to be that given the various components of a document score, having or not having a term, and whether the term is good or bad, the overriding influence on the final score for these collections is not having any query terms, and specifically not having any of many request terms. That this is the case is suggested by the fact that collections like C1400Io, with short requests, suffer only a slight loss of recall even when a sample of only one relevant document is used for weight generation, while the estimation method also works well enough for the two N11500 collections. In other words, the problem is the combination of little information and an inadequate estimation technique. Evidence for this analysis suggesting that it is u rather than w which is causing problems is supplied by a test simply setting u=0 for a term when n=0, but otherwise computing u in the standard way. The results of doing this for variant collections for several of the test bodies of data are given in run set v2/MR2.1 These show some improvement in recall for some long request collections, i.e. for F2 and F3 with the U27000Pbo and E2500Po collections respectively, though not for B2 and B3 with E2500To, while performance for the two short request NPL collections with B2 and B3 is the same as for the original weighting. This is of course an entirely ad hoc procedure, and is designed mainly to check the problem analysis. A more proper, though not wholly satisfactory approach, is to eliminate the term by setting both w=0 and u=0 when n=0. The results of doing this are given in run set v2/MR2.2, for B2 and B3 with E2500To and the two NPL collections; they show no difference from the crude approach. The technique embodied in Harper's weighting formula H1 is, however, clearly superior in principle, since in this case when any of the key elements in the four components of the formula is zero, the whole component is deleted.

It is important in any case to emphasise the fact that even where poor estimation reduces recall, considerable benefits follow from the use

of relevance weighing. This is clearly seen by looking not at the rather abstract recall/precision figures considered hitherto, but at the actual numbers of relevant and non-relevant documents retrieved, which the user is offered for assssment. A detailed analysis of examples is given in Sparck Jones 1979a, so the illustrations given here will be more summary. Figure B2 shows, for the C1400Io and U27000Pbo collections, the average numbers of relevant and non-relevant documents retrieved per request
a) at the lowest matching score, 1, which is the only one plausibly comparable for both terms and weights; and
b) at one low recall level, 30%.
These figures show, in an extremely striking way, how the average number of non-relevant documents retrieved is reduced with relevance weights. In real terms, the loss of some relevant documents is more than balanced by the huge reduction in non-relevant documents, given that for equal scores the user is scanning randomly to extract relevant documents from the set retrieved. For example, for weighting based on FIRST2 for the U27000Pbo collection, the user at a score of 1 trades 18.9 relevant documents amidst 724.2 non-relevant for 12.0 relevant with 146.5 relevant. At recall 30%, he (naturally) retrieves the same number of relevant documents as terms, 7.7, but with 23.9 non-relevant rather than 173.3. (Terms actually retrieve 7.6 relevant, the trivial difference being due to working the figures backwards from the recall and precision values.) Thus for each relevant retrieved the user is scanning an average of 12.2 documents rather than 38.8 at score 1, or 3.1 rather than 22.7 at a recall level of 30%. That is, if the user's interest is in precision, and in reducing scanning effort, relevance weighting even with a defective estimation technique can be very helpful.

The exact way in which relevance weighting achieves these performance figures is indicated in Sparck Jones 1979a by an illustrative analysis of searching with some individual requests. These details are not repeated here, but the main point made is of importance. This is that, as the analysis shows, relevance weighting works in practice by giving a negative score to frequently-occurring terms with low relevant document incidence. As most documents match requests on one or at most two terms, any documents sharing such terms with a request will not be retrieved; and as these documents will typically be non-relevant, just because the terms are frequent, a great many unwanted documents will be eliminated.

Comparing formula F4 with other weighting formulae, and specifically with U1, it is evident that the effectiveness of the latter is due primarily to the 'promotion' of some relevant documents, rather than to the elimination, i.e. demotion below the natural search cutoff, of non-relevant ones.

## 2 Boolean searching

It would be very useful if the type of approach to indexing and searching investigated by the project could be directly compared with more conventional ones. But it is unfortunately the case that it is practically impossible to make meaningful comparisons between searching generating an ordered output, like that described here, and normal Boolean searching providing unordered output. This problem arose in the previous project, since Boolean search results were available for the UKCIS profile data, and it has been discussed by Evans in connection with

the profile data experiments reported in Evans 1975a and b. An additional problem with the UKCIS material, discussed in the 1977 report, is that its relevance assessments are for the Boolean search output and so bias any comparison between Boolean and other search results. The comparisons between weighted and Boolean searching which follow are threfore no more than tentatively indicative: they are intended more as gestures in the direction of the practically important question, how do unconventional and conventional search methods measure up to one another, than as serious attempts to answer this question. The question must be answered, but it needs more, and more appropriate, data than that available to the project.

We therefore simply note that Boolean searching on the U27000Pbo and U27000Pbl collections gives recall 63% with precision 51% and recall 62% with precision 55% respectively, using the given assessment information. However UKCIS own estimate (see Barker, Veal and Wyatt 1974) was that true recall for Boolean searching on titles is 40%. At the same time, as was argued in the 1977 report, we can assume that non-Boolean searches with exhaustive requests must retrieve many unassessed relevant documents. If we further assume that these searches retrieve 90% of all the relevant documents in the collection, and distribute these extra relevant documents proportionally over the matching levels of weighted searching, we obtain improved performance for weighted searching, and specifically better precision for given recall. Thus if the revised Boolean performance figure for U27000Pbl is 40% recall with 55% precision, the revised precision figure for weighted searching at 40% recall is 45%, compared with the original 23%. Moreover since relevance prediction from First to Last is demonstrably inferior to that from Even to Odd, we could expect revised weighting performance for Odd to be very competitive with revised Boolean.

This argument is of course essentially speculative, but it must be emphasised that raw comparisons between the project relevance weighting figures and the original UKCIS Boolean ones are not legitimate, and further, that any legitimate revisions are likely to place relevance weighting in a more favorable light relative to Boolean searching.

Since Boolean searches were also carried out with the Evans data, a comparison between the project results and Evans original ones can also be attempted. However this can again, apart from global methodological considerations, only be a loose comparison, since the detailed output given in Evans 1975a and b is for profile and document subsets. Using the data given in Appendices 9 and 11 of Evans 1975a, and specifically adjusting the total relevant given in the latter to allow for the smaller number of profiles of the former, we obtain an average Boolean performance for 30 profiles of 34.4% recall and 54.7% precision. This compares with 28% precision at 40% recall and 38% precision at 30% recall for relevance weighting by F4p on the E2500Po collection, having more requests and the same number of documents.

It thus appears quite difficult to achieve levels of performance with unconventional approaches as good as those apparently obtained with conventional Boolean searching; but we cannot draw any firm conclusions from such very weak detailed comparisons. Much more work needs to be done on this whole problem.

At the same time, the superiority of system-derived relevance weights over the user's intuitive request term weighting is suggested by a comparison between the two which is possible for the E2500Po collection. The search results for the user weights, labelled 'Human wts', are given in Other Table run 02. F4p is superior to the user's own weights, which perform about the same as collection frequency weights and F1p. (It is of interest that in Miller's 1970 experiments, relevance weighting was based on F1, and these were calculated using (simulated) user estimates of request term value.)

## 3 Comparable research

The project research is related to some done by other projects, and, specifically, must be compared here with experimental work done elsewhere.

The obvious comparison relating to request membership is with Harper and van Rijsbergen's tests, already mentioned in Section B; the important comparisons relating to relevance weighting are with the various experiments carried out by Harper, Porter, van Rijsbergen and Robertson. Harper et al. have indeed, as indicated earlier, taken request membership and relevance weighting together, so their work is most conveniently considered without an explicit division between the two factors. The results they have obtained are first described, and are then related to the project findings.

Harper and van Rijsbergen 1978 reports tests of the term dependence model using the C1400I collection and the recall cutoff representation method. In a series of upper bound experiments (applying predictive formulae retrospectively) they compare terms and independence weighting by F4pr, combined with MST expansion, with the strict dependence model involving both weighting and expansion. We can refer to the former in a style linking these tests with our own as F4pr/MST, the latter as D/E. (Harper and van Rijsbergen use I for F4 and E for expansion by either means.) As controls these are compared with independence weighting of the original query without expansion, i.e. F4pr/Q, and with the dependence weights without expansion, labelled D/Q. The comparisons show T<<F4pr/Q=F4pr/MST=D/Q<<D/E: i.e. that it is the combination of dependency weighting with expansion in the pure model which is most effective. Harper and van Rijsbergen then establish that the Harper weighting formula H1 plus MST expansion as used earlier, which we may label H1/MST, is as effective as the strict dependency model, i.e. H1/MST=D/E, at any rate retrospectively.

These tests establish the improvements attainable in principle; the second group of experiments reported tested feedback, i.e. iterative searching, using the top-ranking 10 or 20 documents obtained by a coordination search to compute weights for searching the remainder of the collection. Performance is given for the latter only, the residual collection. (In these tests 0.5 is used with H1.) For this environment Harper and van Rijsbergen show that term frequency information other than that referring to relevant documents should be taken from the unsearched (large) collection rather than from the sample. (that refering to relevant document occurrences has of course to come from the sample.) Calling this version of the procedure H1/MST*, Harper and van Rijsbergen then show that T<<F4p/Q<=H1p/MST*<<H1pr/MST*, i.e. that the dependency

model in the heuristic form represented by H1/MST is more effective than the simple independence model but, interestingly, not strikingly so. It should be noted that F4p/E was not included in these comparisons, but presumably as it performs the same retrospectively as F4pr/Q, it could not be expected to perform relatively better predictively. In these runs requests having exhausted all their relevant documents in the top 10 or 20, or not having retrieved any relevant, were excluded from the second search.

Van Rijsbergen, Harper and Porter (in press) report further experiments with more collections, i.e. with the U27000T CAC sets (i.e. First and Last, U27000Tf and U27000Tl) as well as with C1400I. The tests were with feedback as before, and compared the heuristic dependency model, i.e. H1/MST, H1 weights plus enlargement, with F4/MST, Formula F4 plus enlargement, obtaining the result for a sample of 10 that T<F4p/MST<H1p/MST*, and for a sample of 20 that T<<F4p/MST<H1p/MST*. They then compare a variety of alternative association measures as a basis for the formation of the MST classification, but find little consistent or real difference.

A most interesting analysis shows that for specific rank cutoff points in the residual collection documents retrieved by the second search, the weighted and enlarged feedback with H1/MST* retrieves far more relevant documents than the simple term search, and manages to retrieve at least one relevant for a larger proportion of the requests.

It should be noted that in these tests, unlike the previous ones, 0.5 was not used as an estimator as it was recognised as unnecessary.

Robertson, van Rijsbergen and Porter (in press) is devoted mainly to the development of a comprehensive probabilistic weighting and searching theory, combining relevance weighting of the type dealt with in this report with Harter's Poisson distribution approach to term selection exploiting within-document frequency information. The experiments reported were carried out with the N11500A collection, in fact used predictively both from Even to Odd and from Odd to Even. Broadly speaking the results for the two are the same.

The results of interest are those comparing the techniques studied by the present project with these techniques as it were extended by the addition of Harter Poisson-based weighting emphasising within-document term frequency information. Thus collection frequency and Harter weights are combined to form what may be referred to as P0, and Harter weights and relevance weighting are combined to form what may be labelled P1 (this is not simply an addition or multiplication of the two, but is slightly more complex). Thus the experiments compare on the one hand simple collection frequency weights using F0 (treated as a special case of F4 as suggested by Croft and Harper) and relevance weighting using F4 on the one hand, with P0 and P1 utilising additional information on the other. The tests are based a) on the use of all the relevance documents in the weight generation collection and b) on the top ranking 20 (which will be identified by /20). The predictive experiments are complemented by a retrospective comparison between F4pr and P1pr. Estimation in P1 is done in a manner resembling the use of 0.5 for F4.

We thus compare terms T with F0, F4p, and F4p/20, with P0, P1p and P1p/20, and with F4pr and P1pr. The result is

T<<F0=P0=F4p/20=P1p/20<=P1p<=F4p<<F4pr=P1pr. There thus appears to be no particular advantage in the Harter Poisson weighting; but at the same time relevance weighting with poor samples is no better than collection frequency weighting. It should be noted that the results for F4p/20 are similar to those obtained by the present project for F4p B3.

The work just described is the most extensive on relevance weighting requiring comparison with the present project tests. Other work on relevance weighting parallel to that done by the project has been carried out chiefly by Salton and Yu and their colleagues. The earlier work done at UKCIS was discussed in Sparck Jones and Bates, and it is sufficient to repeat here the main result of Robson and Longman's tests (Robson and Longman 1976), namely that with profiles obtained by iterated searches of one document set exploiting formula U1 to order and hence select relevant document terms, performance on a second set was quite competitive with that given by carefully honed manual profiles: precision was 24.1% and recall 57.8%, comapred with 36.3% and 76.5%. This research does not appear to have been pursued further, which is to be regretted given the opportunities provided by the UKCIS service for testing in an operational environment.

The early SMART work on relevance weighting was discussed in Sparck Jones and Bates. More recent papers, especially Yu and Salton 1977 and Yu, Lam and Salton 1980, have been devoted to theory. The latter in particular shows that under certain conditions relevance weighting, referred to as term precision weighting, offers the best possible term weighting system. The paper relates relevance weighting using F4 to the SMART discrimination weighting formula ("Q") which has both within-document and collection frequency components, i.e. represents a development analogous to that in Robertson, van Rijsbergen and Harper's paper. Unfortunately there have not yet been any significant tests of the SMART ideas, especially in a predictive context.

Vernimb 1977 describes an interesting application of relevance weighting within an operational context (the ENDS system). The motivation is somewhat similar to that of Barker,Veal and Wyatt's 1972 study, namely relevance weighting is exploited as a device to help the user, in this case in his assessment of the output at each phase of his search interaction. Thus the documents retrieved by a Boolean search are ranked by a relevance weighting formula for presentation to the user. The ranking hopefully concentrates the relevant documents at the top of the list, so promoting effective request reformulation either by the user or as part of an elaborate automatic reformulation procedure utilising the term indexing of the relevant documents, i.e. as part of a kind of Boolean request substitution process. The use of relevance weighting is described as routine for the first purpose; the paper presents examples of the more complex automatic procedure, and Vernimb maintains that this procedure, while not necesarily improving recall, will improve precision. It is extremely unfortunate that systematic and controlled experiments, especially designed to investigate the relative contributions of relevance weighting and the other components of the request substitution process (which includes the use of simple collection frequency weighting obtaining w = 1/n) do not appear to have been carried out. If we define the output of a search as a collection, i.e. define N as the output size and n relative to this, Vernimb's formula, which we may label V1, is

$$w = r/R + ((n-r)/(N-R)).$$

## 4 Conclusion

In the account of the experiments in Section B the only methods of performance representation mentioned were document value and recall cutoff. Further views of the behaviour of relevance weighting are provided by the other secondary methods, though apart from the total retrieved figures of the Str Tables, they are available for only a few searches. The way in which relevance weighting eliminates non-relevant documents is clearly shown by the Str total retrieved Tables. Some examples for F4p are given below: in all of these cases, using the full relevance information in the weight generation set, the same or similar numbers of relevant documents are retrieved using F4p as for terms T, but there are substantial reductions in the numbers of non-relevant.

AVERAGE NON-RELEVANT RETRIEVED

| | T | F4p |
|---|---|---|
| C1400Io | 378.7 | 66.9 |
| U27000Pbo | 743.1 | 354.1 |
| N11500Ao | 1584.9 | 347.1 |
| N11500To | 666.3 | 408.2 |
| E2500To | 377.0 | 133.1 |
| E2500Po | 240.4 | 91.7 |

The promotional effect of relevance weighting is shown by the Scr Tables, giving the proportion of requests retrieving their first relevant by a given rank (in fully-ordered output). The examples given below, illustrating the same comparisons as the previous one, show that for all of the collections except E2500Po, where term performance is unusually good, relevance weighting increases the chance of getting a relevant document at high rank.

PROPORTION OF REQUESTS RETRIEVING FIRST RELEVANT BY RANK 10

| | T | F4p |
|---|---|---|
| C1400Io | 73.8 | 82.9 |
| U27000Pbo | 83.3 | 90.3 |
| N11500Ao | 80.9 | 92.1 |
| N11500To | 76.4 | 84.3 |
| E2500To | 84.2 | 92.1 |
| E2500Po | 94.7 | 94.7 |

Unfortunately, as mentioned in Section A, some of the alternative performance representation methods are rather costly. The main justifications for the use of document value are that it is cheap, and that comparability with earlier results is maintained. That for recall

cutoff is comparability with other projects, especially Smart. A good case can nevertheless be made for document rank, and further work needs to be done on what view of performance this provides for such searches as those on the U27000Pbo collection with weighting formula U1, for which divergent evaluations are provided by the document value and recall cutoff methods.

Attempting now to assess the project results overall, we can list, in the form of questions, the propositions the project was intended to verify. Thus referring back to the factor characterisation of retrieval systems provided in Section A, with respect to indexing variables we ask:

A1) Is relevance weighting appreciably more effective than simple term matching or weighting without relevance weighting?

A2) Is relevance weighting using the recommended formula F4 most effective?

A3) Is relevance weighting made more effective by request enlargement, specifically that using relevance information to select extra terms?

With respect to input factors, treated by the project as environmental parameters rather than system variables, we ask:

B1) Is relevance weighting effective for different indexing modes?

B2) Is relevance weighting effective for different indexing sources?

B3) Is relevance weighting effective for different indexing description exhaustivities?

With respect to output factors, especially output variables, we ask:

C1) Is relevance weighting effective as an alternative to the conventional Boolean method of controlling matching?

These questions concern explicit or implicit system variables. There are also questions referring to the influence on relevance weighting of broader environmental parameter settings. Thus we ask:

D1) Is relevance weighting effective for different subject areas?

D2) Is relevance weighting effective for inhomogeneous collections?

D3) Is relevance weighting effective for poor term occurrence information?

D4) Is relevance weighting effective for slight relevance information?

We have also to ask a specifically methodological question, namely:

E1) Is relevance weighting effective for different methods of performance representation?

And we should perhaps ask a final question:

F1) Is relevance weighting effective in any particular way?

Given the variety of results obtained in relation to some of these questions, and the limited range of results relating to others, the project answers to the questions must in some cases be rather crude generalisations. However the testing was comprehensive enough to justify definite statements in some of the most important cases.

The project answers to the questions are:

A1: Relevance weighting is typically more effective, and indeed very much more effective, than simple term matching or weighting without relevance information.

A2: Relevance weighting by F4 is typically more effective, and often conspicuously more effective, than other weighting based on other formulae.

A3: Relevance weighting may be made more effective by being combined with request enlargement, even by crude methods

B1: Relevance weighting is apparently not seriously affected by indexing mode.

B2: Relevance weighting is apparently not seriously affected by indexing source.

(It must be emphasised that the answers to B1 and B2 assume no concomitant differences in other variables like exhaustivity.)

B3: Relevance weighting appears to lead to greater performance improvements with longer request or document descriptions, especially the former.

C1: (On the extremely poor evidence available) relevance weighting performance can approach, but not exceed, that of Boolean searching.

D1: Relevance weighting is effective for different subject areas.

D2: Relevance weighting is effective even for inhomogenous document sets.

D3: Relevance weighting is quite effective even with poor term distribution information.

D4: Relevance weighting is effective as a precision device even with slight relevance information.

E1: Relevance weighting is effective for different methods of performance representation, but variably.

F1: Relevance weighting is a precision device.

These answers refer in particular to weighting by F4, since this proved to be generally most effective. However, as the test results show, F4 is very sensitive to the estimation technique used, and it is clear that a better one must be found if question D4 is to be given a less qualified answer. Equally, the behaviour of U1 in those cases where F4 is most afflicted by estimation problems suggests that U1 is effectively a quite robust approximation to F4.

The project tests thus support some positive answers to the questions asked about relevance weighting. But it is also evident that

the experiments, though quite extensive, do not answer all the questions sufficiently definitely. In particular they do not, chiefly through the lack of adequate data, provide a clear enough view of the detailed effects on relevance weighting of the input variable values, especially for source and exhaustivity, and of environment relevance parameter settings. The tests further bring out very sharply the persistent challenge of performance representation.

The way to start tackling these problems is nevertheless quite clear: with more data, adequate evidence about input variables and parameter settings, and also about the relative merits of weighted and Boolean searching can be obtained, using the the experimental apparatus exploited in this and the previous project. Those questions which might be regarded as newly raised by the project, namely about the value of crude request enlargement techniques when combined with weighting, and about the value of the crude weighting technique represented by U1, could also be tackled with this apparatus. The methodological problems of performance representation could also be studied using the range of programs available, since the project failures here were due to lack of time rather than lack of test apparatus.

In other words, a natural way to try to consolidate the answers already given to questions about relevance weighting, and to provide new ones, is to crank the handle on the production engine already running in Cambridge.