Section A : the 'ideal' test collection and obtaining relevance assessments for it

A1 Introduction: the 'ideal' information retrieval test collection

In 1975 it was suggested that a high-quality, general-purpose information retrieval collection, the 'ideal' collection, would be a material aid to information retrieval research (see Sparck Jones and van Rijsbergen (1975)). In a subsequent design study project, a detailed specification for 'ideal' collections meeting a range of requirements was worked out and costed (Sparck Jones and Bates (1977)). An 'ideal' collection would essentially consist of at least one set of documents indexed in various ways, with an associated set of requests, also variously indexed, and relevance judgements on the document set for these requests. The 'ideal' collection, which would be derived largely from operational services, would be sufficiently well-founded, and sufficiently fully characterised, to support a great variety of well-organised and methodologically proper experiments.

The most intractable problem of the 'ideal' collection is the provision of relevance assessments. This is because the document set, if it is to meet some important experimental needs, has to be large, say of order 30K, which is too large for exhaustive assessment, i.e. the evaluation of every document in relation to every request, at any rate on any realistic costing. Partial relevance assessment is not necessarily a problem for restricted experiment: calculating recall for say either of two indexing or searching strategies in relation to their pooled search output may be quite in order. But the 'ideal' collection would be set up for open-ended use. That is, relevance information has to be supplied when the collection is built which is adequate for any future experiments evaluating wholly novel strategies. (Assessment of new search output subsequent to building is theoretically possible, but is methodologically unsound and practically inconvenient.)

The crucial question is therefore whether assessments can be provided, at the time the 'ideal' collection is built, in such a way that in virtually any future experiments using the collection, valid performance comparisons between indexing or searching strategies can be made. This is primarily a statistical question. An initial attack was made on this question in the design study report (Sparck Jones and Bates (1977)), but the present project was intended to

investigate it in much more detail. However since assessment is expensive if done on a large scale, the question is really whether assessments can be provided in a statistically-valid but not intolerably expensive way.

A2 The 'ideal' collection specification

The specification of the 'ideal' collection given in the design study report (Sparck Jones and Bates (1977)) is quite complex: alternative collections with different costs are presented, and any one collection may in fact consist of more than one document set with requests and relevance assessments. Further, the characterisation of the collections was in part determined by the method of establishing relevance assessment requirements worked out as part of the study. This, the 'Pool' method, is considered in detail in Section B. For present purposes it is sufficient to note that it proposes the assessment of a random sample of the comprehensive pooled output for a range of alternative searches for a given request, and that as the number of documents to be assessed per request varies inversely with the number of requests, a large request set is desirable.

As part of the present project work was an investigation of the implications for collection building of any proposed methods of determining assessment requirements, a summary of the 'ideal' collection specification is given here for reference. This is most conveniently provided as a description of the option C 'ideal' collection, representing the middle collection in a range of five. Option C, costing from £80K depending on data pricing, would be as follows:

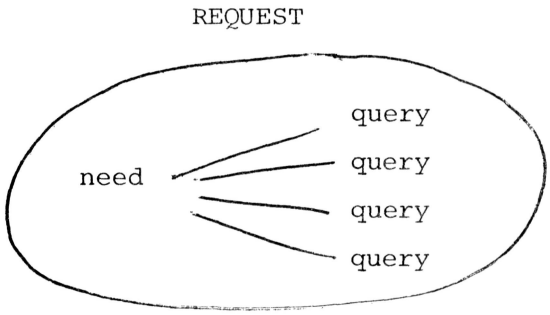
documents	main set	30K, in science (e.g. from Inspec)
		having titles, abstracts and three other forms of indexing, namely keywords, thesaurus descriptions, and high-level subject codes
	other set	3K, in social science/humanities
		comparable in size with a random subset of the main set; indexed in the same style as the main set.

The main and other sets would have core characterisations, representing the information just described, plus citations; some subsets of the main set, and specifically the random subset, would have enriched characterisations in the form of further indexing.

requests — As the material would be obtained from an operational service, and specifically from on-line searching, a request is not a simple entity. It is defined as a complex whole consisting of

a) the user's original need statement
b) the query form(s) in which it has been searched for him
c) alternative query formulations representing systematic uses of different indexing options
d) any further query formulations used as methods of increasing the output set required for evaluation.

Thus the request may be represented thus:



primary set — ~700, for the main set of documents

secondary set — ~200, for the other set of documents

The core characterisation of requests would consist of the need statement a) above, and the systematic queries under c) (along with any effectively systematic queries under b)); enriched characterisations for request subsets would represent further systematic alternatives.

relevance assessments — The 'Pool' method, related to observed output from 30K document sets, suggested that the 10% random sample from a comprehensive pool would require an average of about 300 assessments per request; these would be primarily by the user and would be graded.

A3 Relevance data

One practical influence on the statistical study, namely the cost of implementing consequent assessment procedures, has already been mentioned. Another is the actual facts of document retrieval systems, and specifically the characteristics of real relevance data. Thus, to take an imaginary example, if a statistical method required a minimum of 100 relevant documents per request, where many actual requests have, or may be presumed to have, far fewer, even in a large document set, this would have significant implications for the representativeness of the 'ideal' collection. Professor Cleverdon raised points of this kind in connection with the application of the 'Pool' method in the design study.

To provide a background for the statistical study in relation to such considerations, some facts about the relevance properties of actual collections are given here. These are some of the larger collections used in different experimental investigations, for which data is available. Thus for the Cranfield 1400 collection, Evans 2532 collection, UKCIS 27361 collection, NPL 11571 collection

The table below shows

the number of documents in the test set
the number of requests
the total number of relevance postings
the number of different documents represented in the total postings and percentage of the whole document set
the number of documents in the shortest request relevance set
the number in the longest set
the mean number of relevant documents per request
the variance
the standard deviation.

	Cranfield 1400 *	Evans 2532	UKCIS 27361	NPL 11571	Lancaster Medlars 1968
no docs	1400	2532	27361	11571	~600000
no reqs	225	39	182	93	302
no rel post	1614	899	10715	2090	na
no rel docs	831	633	7883	na	na
% docs	59.4	25.0	28.8	na	na
min rel	1	3	1	1	na
max rel	40	53	554	84	na
mean rel	7.2	23.1	58.9	22.5	85.7**
var	29.2	213.7	na	328.3	na
st dev	5.4	14.8	na	18.2	na

* exhaustive assessments

** estimated: there were 9.7 relevant in 19.7 assessed, on average, i.e. 49.2% relevant; as the assessed were a random sample of the retrieved this implies 49.2% of the retrieved 174.1 were relevant, i.e. 85.7.

In Appendix 1, Figure 2, histograms showing the actual distribution of relevant document sets are given for some illustative collections for which the data is available.

A4 Constraints on statistical methods

It will be evident from the table above that request relevance sets show very great heterogeneity. We may therefore list, as factors to be taken into

account in considering methods for determining relevance assessment requirements, some points as follows:

a) relevance data heterogeneity as a whole
b) very long lists
c) very short lists

Such points are clearly important when any kind of sampling is envisaged. Further, underlying them is a range of retrieval system characteristics which have to be borne in mind in considering the applicability of statistical methods, though they do not bear directly on the formal characterisation of the methods themselves. These are:

i) request specificity variation
ii) user need variation
iii) document set heterogeneity
iv) relation of document subsets of, say, size 30K, to large files of, say 300K.

These factors are related to what may be called the relevance assessment input to the 'ideal' collection. Another class of factor is associated with its output, i.e. use in future experiments. The 'ideal' collection is specifically intended to be hospitable in the sense of providing adequate test data for new lines of research, largely because in the past collections made for one research purpose have been found very inadequate for others. The 'Pool' method as presented in the design study report requires a comprehensive search output pool as the basis for assessment for each request, i.e. a pool containing all the relevant documents for the request, and all the output of plausible future searches for the given need statement, i.e. extensions of its query set. When more exotic search strategies are considered, e.g. some types of document clustering, it is clear that these requirements are unrealistic, and less exigent statistical methods are therefore needed. Further, the 'Pool' method assumes evaluation of retrieval strategy effectiveness in terms of recall or precision, or their relatives. An important question is thus how much restriction can be accepted in evaluation in any future experiments. Additional

factors to be considered in the choice of method are therefore:

α) assessment scope and reliability

β) assessment utility.