Section C : practical implications of the statistical methods for building and using the 'ideal' collection

C1 Summary of the properties of the statistical methods

Section B provides two bases for obtaining the relevance assessments of the 'ideal' collection, the Pool method and the Squares method. Each relies on certain assumptions and has certain implications in relation to the real properties of any collection data, with in turn consequences for the practicalities and costs of building the collection, and in relation to the properties of strategies to be tested in experiments with the collection. The reader is reminded that while the whole basis for the discussion is the requirement that adequate data should be provided for the comparative evaluation of future indexing and searching strategies, it is recognised that neither the whole output of searches used to build the collection, nor that of future searches, can be assessed. Thus from some points of view the relevance information supplied with the 'ideal' collection could be inadequate through being incomplete. In both human and economic terms we cannot expect full relevance status information, though additional information could of course be gathered for specific projects. This being the case the object of the exercise has been to establish methods for obtaining enough information for statistically reliable strategy comparisons, and also a fair range of investigations requiring access to specific documents of known relevance status.

However, some constraints have to be placed on the form of such comparative evaluation, which are expressed as assumptions underlying the methods. Thus it is assumed (assumptions 1.1 and 1.2) that performance evaluation depends on recall and either precision or fallout, and also that these can be estimated by proportions based on samples. This is not to say that other forms of evaluation, for example simple numerical ones, are precluded: any associated with either the number of documents of known status, or the simple totals of matching documents, are clearly allowed. Quite other approaches to system evaluation are not allowed for, but the original design study for the 'ideal' collection did not suggest any other approaches more generally acceptable than those based on recall and precision/fallout.

The other assumptions applicable to future experiments with the collection, namely 1.3-6, concern the statistical properties of the evaluation data, i.e.

data about the strategies being compared. These assumptions were considered in Section B, and especially in chapter B5. In general the respective assumptions made for the two methods are not extreme, and where specificance level or numerical values have to be supplied as, for example, for the power of the test used, the main implication of more exigent values is to increase the number of assessments required, by relatively moderate percentages.

The assumptions made about the assessment data used to obtain the relevance information are the same for the Pool and Squares methods. They include two, 2.3 and, 2.4, about the properties of samples from the assessment pool, which are not very stringent; and two, 2.1 and 2.2 about the content of the pool, which are stronger but which can be abandoned, though the pool size must then be known. They may alternatively be weakened for the Pool method, which increases the number of assessments required by a relatively small percentage.

The assumptions about the request data, in relation to both assessment and evaluation, include the common one, 3.1, that the requests are independent, which may be accepted for a large request sample (and is indeed usually assumed in retrieval experimentation). The strong assumption about request homogeneity needed for the Pool method is not needed for the Squares method, an advantage of the latter which has already been mentioned.

Thus summarising the really distinctive requirements of the methods we have:

<u>Pool</u>

the pool contains all relevant (or output) documents
the requests behave the same

<u>Pool, weakened</u>

the pool contains a specified percentage of relevant (output) documents
the requests behave the same

<u>Pool, modified</u>

the pool size is known
the requests behave the same

<u>Squares</u>

the pool size is known

Further, the Pool and Squares methods are globally distinguished by the fact that the former refers to individual requests, the latter to a set of requests:

thus the pool for the former is the pool for each request, where the pool for the latter is in fact the sum of the individual request pools. Individual requests only enter into the Squares method as the means of providing the totalled information.

C2 Application of the methods

C2.1 Testing to obtain statistical method parameter values

It is evident that applying either Pool or Squares method in building the 'ideal' collection must involve a preliminary sampling of whatever services are used to provide the collection material, to gather information about request properties, e.g. the number of relevant documents they have, the number of documents they retrieve by the standard strategies, the way in which the output can be extended to obtain a pool, the extent to which the pool is exhaustive, etc. Such testing of the service used would of course add a cost to the building operation, but as, for example, individual document evaluation need not be done very carefully, the cost would be small. Thus £1000, say, would cover a good deal of sampling.

This testing of the service would produce the appropriate numerical parameters for the methods, and consequently the number of requests required for the Pool method, or a prediction of the order of number required to supply n, the total of known documents, for the Squares method.

C2.2 Practical implications of the methods

Clearly, to determine the practical implications of either method, in terms of assessment effort, we should consider the least, or at any rate less, favourable cases: for the data little performance difference between the strategies, few relevant documents per request, for the test fairly high power in the test, etc. Considering the absolutely worst cases implies very extensive assessment, but this is perhaps unrealistically gloomy: however a relatively conservative approach is called for in making the explicit comparison between the methods which is necessary here. The two methods were compared in chapter B5, but a more detailed treatment is required as a basis for recommendations as to the proper procedure to be adopted in building the 'ideal' collection.

In spelling out the implications of the Pool method, we assume that the difference between the two strategies A and B being investigated in the future is not very large, i.e. $p_A - p_B = 5\%$, and that 5% significance and 95% power in the test are needed, which is fairly conservative. We now assume that the requests have an average of 25 relevant documents, less than half the average for the UKCIS data described in Section A: we consider recall as its requirements are more stringent in real terms than those for precision or fallout. Referring to Table 1, 25 relevant documents per request implies assessing 20% of the search pool obtained when building the collection, for each of 1000 requests, and 36% for 500 requests: this would give evaluation samples of 5 and 9 relevant documents respectively in the outputs of A and B. Since there is some doubt about the feasibility of getting 1000 requests, or the convenience of such a large set for future experiments, we consider 500 requests. If the pool has an average size of 3000 documents, which is somewhat larger than that obtained with fairly hospitable searches of the UKCIS data, this gives an average of 1080 assessments per request, which is a lot. If assumptions 2.1 and 2.2 are weakened so only 90% of the relevant documents are expected in the pool, the number of assessments is increased from 36% to 40%, which implies 1200 assessments from a pool of 3000. However in this case a pool of 2000 might be realistic, implying 800 assessments.

If the modified Pool method is used, where assumptions 2.1 and 2.2 are abandoned, but the pool size must be known, Table 3 supplies figures as follows. If it is assumed that there are 25 relevant documents for the request in the pool, and an evaluation sample of 9 relevant is required, and the pool contains 100 documents, then 49 of these must be assessed. This is perhaps a rather unrealistic pool: if we assume more realistically that there are 25 relevant in a pool of 500 documents, then 250 would have to be assessed. If there are 1000 in the pool, we require 502 assessments. Such figures apply to individual requests, but if we also treat them as averages, and compare the original and modified Pool methods for the 500 requests for which an evaluation sample of 9 is required, and assume a pool of 1000 documents in both cases, the original method requires 360 assessments compared with the modified method total of 502. However this is not quite a proper comparison since the original method requires a comprehensive pool: thus if 25 relevant documents can be retrieved in a balanced but non-comprehensive pool of 1000 a corresponding comprehensive pool for, say, a larger inclusive set of 30 documents, might well have to contain 1500 documents, and then 540 of these would have to be assessed. It should be noted that we wish to be confident that the required evaluation sample is included in the assessed set: the modified Pool method requires more assessments than the original, but in compensation the probability that the required number of evaluation documents is in fact achieved increases from 0.50 to 0.95.

Turning now to the Squares method, we again assume 5% significance and 95% power, and analogously assume that the difference between strategies A and B is not large, taking $p_b + p_c$ (i.e. Π_R) = 0.25: this means that the percentage of documents treated differently by the two strategies is low. We than look for a total of retrieved documents, n, large enough to keep $\frac{p_b}{p_b + p_c}$ (i.e. Δ) small as this reflects lower expectations about the characteristics of the strategy outputs. Thus, refer to Table 7, if we take $\Delta = 0.553$ as the value of Δ associated with a lower confidence bound figure for b + c, i.e. the least number of documents to be retrieved by either strategy A or B alone, this implies n = 5000. (b + c = 1190 and strategy A is superior to B if b, the number of documents retrieved by A alone > 629.) Thus in the case of recall we are looking for 5000 relevant documents. If we then assume that there is an average of 25 relevant documents per request in a pool of 1000 documents (this need not of course be all the relevant documents for the request), this implies a set of 200 requests. However the 25 relevant documents would in this case only be obtained by exhaustive assessment of the pool. If we consider, more realistically, the assessment of half the pool, i.e. 500 documents, we need 400 requests. If we assume the same level of sampling as for the Pool method, namely 36%, we would expect to get 9 relevant documents for 360 assessed, and hence would have to have 556 requests. However we note that a slightly larger $\Delta = 0.568$ is associated with n = 3000. which 36% sampling of a pool of 1000 documents implies 334 requests. It must be emphasised, however, that these are crudely obtained figures in that we cannot rely on getting the required number of relevant documents in the assessed sample more than 50% of the time. If we wish to be

95% confident of getting them this implies, for n = 5000, that we should, if we only sample 350 documents, need substantially more requests (alternatively we must use a larger sample).

These analyses suggest that in formally comparable situations, i.e. for the same size of pool, the methods do not differ greatly: the important difference is in the expected size of pool for the requirements it has to meet. Thus the original Pool method requires an exhaustive pool, which is liable to be large and hence to generate more assessments. Neither the weakened nor modified Pool methods require such a pool, so we can expect a smaller pool and hence fewer assessments. In terms of assessment effort these methods then look very comparable with the Squares method. The situation is summarised in the table below.

	pool	no. requests	eval. sample	%assessment	no. assessed	
Pool method	3000	500	9	36%	1080	
	2000	500	9	36%	720	(1)
	1000	500	9	36%	360	(2)
Pool method, weakened i.e. 90% relevant in pool	2000	500	9	40%	800	
Pool method, modified	2000	500	9	∿ 50%	∿ 1000 est.	(3)
	1000	500	9	50%	502	
Squares method	1000	556	-	36%	360	(4)

Comments

(1) rather low size of exhaustive pool

(2) very low size of exhaustive pool

(3) rather high size of non-exhaustive pool

(4) rather hopeful about the sampling: more requests are perhaps needed, though the assumption about Π_R is rather pessimistic.

C2.3 Assessment of the methods

The previous paragraph shows that from the assessment point of view there need not be any large difference between the methods in terms of the practical effort required to implement them in building the 'ideal' collection. Thus the general level of costing given in the design study report is sufficiently accurate. The specific costs will depend on the actual characteristics of the documents and requests of the operational service used to supply the collection data; and these can only be established by the sampling mentioned in paragraph 2.1, supported by the experience of the service operators. We may base some generalisations on such information as has been gathered in the past by those conducting operational system evaluations: so, for example, it would be unreasonable to expect 100 relevant documents on average in a collection of 30,000; but it would equally be a mistake to expect no more than 5.

The choice of technique for obtaining the relevance information for the 'ideal' collection must therefore be based on an overall, broad, rather than narrow economic, view of the methods proposed; that is, on what features of the Pool method, or the weakened or modified versions of it, or of the Squares method, are particularly welcome or unwelcome: the choice for the 'ideal' collection building depends on the weight we attach to these features.

To consider the Pool method. In this we can confine ourselves to the modified version since this is clearly preferable to the original or weakened versions in imposing less stringent requirements on the pool. At the same time, though it requires more information, namely the actual size of the pool for each request, this is not a significant practical problem: the pool size could almost certainly be easily obtained, and simple reference to a preconstructed table would indicate the appropriate percentage for assessment. The Pool method has the advantage that it supplies information for the evaluation of future strategies in terms of recall and precision, each of which, and the latter especially, are felt to be important in relation to operational systems. As Table 1 makes plain, assessment information adequate for recall would be adequate for precision as well, since recall-oriented assessment for 25 relevant documents, say, could be expected to cover a precision-oriented assessment for 25 retrieved documents. The second advantage of the method is that it imposes no requirements on the conduct of future experiments. The disadvantages of the method are first, and most importantly, that it is assumed that the requests all behave in the same way. This strictly implies that the evaluation sample is a minimum and not an average, so that if it is treated as an average, some uncertainty is carried over to the evaluation of future strategies. In this connection, any requests not having the supposed number of relevant or retrieved documents in the assessment sample and hence an inadequate evaluation sample present a problem, which is not properly dealt with by eliminating them from the request set, as this biasses the set. Some sort of lashup is inevitable here such as, say, leaving out only the few worst offenders, and accepting some deficiency in the evaluation samples for the others. The second disadvantage of the method is that it is impossible to use subsets of the query set for more economical future experiments without undermining the conclusions to be drawn from them.

With respect to the Squares method, its merits are first, that it avoids the presupposition that the requests are homogeneous, and so is hospitable to requests with few relevant documents, and to small pool samples for assessment. In consequence, for any real data, it provides for more reliable assessment for future strategies. The method's second merit is that it permits the use of request subsets in future experiments, subject to the implications, for their evaluation, of smaller n: as Table 7 shows, the smaller the total of relevant or non-relevant documents, the more difficult it is to establish strategy differences; however relatively specific statistical statements can be made about the comparative status of the two strategies being compared, which is not feasible for request subsets with the Pool method. The disadvantages of the Squares method are first, that performance evaluation in terms of precision is not allowed: and whatever the theoretical merits of fallout, precision is felt, as mentioned above, to be an important property of real system performance. The second disadvantage is that in any future experiments, specific information about the distribution of retrieved documents has to be obtained; i.e. to fill the contingency table cells we have to know not merely how many documents are retrieved by strategies, but which strategies retrieved which documents: and this may be practically very inconvenient. It should also be noted that unless a reasonable pool sample is assessed, there may be very little information about relevant documents for some requests, which could be a nuisance in other types of investigation, e.g. into the characteristics of relevant documents.

On the whole, the conclusion we draw is that we cannot recommend one particular method unreservedly as the only proper one to use for building the 'ideal' collection: both the modified Pool method and the Squares method have advantages and disadvantages. On the other hand, either, sensibly implemented, would be sufficiently satisfactory. We therefore recommend that the decision be taken by the collection builder, on the basis of a summary poll of likely collection users.

Note

Whatever the application of the arguments of this report to building the 'ideal' collection, we would like to draw the reader's attention to Table 2 taken in conjunction with the data of chapter B2.This table gives, within the context of the Pool approach, figures for the relevance information which is needed, in the case where perfect information can be obtained, to evaluate strategies tested for particular request set sizes. The figures reinforce the observation that many tests in the past could have been inadequate, in the sense that the data did not statistically support the conclusions drawn from it; or at any rate that, if they were not inadequate this is because the data had additional, statistically relevant properties, which were nevertheless not identified or indicated.