Using Interdocument Similarity Information in Document Retrieval Systems

Alan Griffiths, H. Claire Luckhurst, and Peter Willett*
Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

The first part of this paper reports a comparative study of the document classifications produced by the use of the single linkage, complete linkage, group average, and Ward clustering methods. Studies of cluster membership and of the effectiveness of cluster searches support previous findings that suggest that the single linkage classifications are rather different from those produced by the other three methods. These latter methods all produce large numbers of small clusters containing just pairs of documents. This finding motivates the work reported in the second part of the paper, which considers the use of clusters consisting of a document together with that document with which it is most similar. A comparison of the use of such clusters with conventional best match searches using seven document test collections suggests that the two types of search are of comparable effectiveness, but they retrieve noticeably different sets of relevant documents.

Introduction

A central problem in document retrieval is the identification of a few relevant documents that can form the basis for a relevance feedback search using probabilistic retrieval methods [1,2]. The simplest means of identifying such documents is to carry out a full search in which some matching function is used to determine the degree of similarity between each of the documents and the query. The calculated match values may then be used to identify the most similar documents, and thus those that are most likely to be relevant to the query [3]. A rather more sophisticated approach attempts to use approximate probabilistic models that do not involve the use of exact relevance information [4]. A third approach, and the one discussed in this paper, involves the use of clusters, or groups, of documents in which a query is matched against the clusters, rather than against individual documents [5-9]. Such a retrieval strategy may be more effective than a conventional search since the relationships between documents are taken into account when deciding which documents are to be retrieved, as well as the relationship between the query and the individual documents.

A wide range of clustering methods has been suggested for the grouping of documents in bibliographic retrieval systems [10]. Of these, the most effective would seem to be methods that are based on a similarity matrix that contains the similarities between all pairs of documents in a collection. Typical of such procedures are the hierarchic clustering methods that operate by means of a series of agglomerations in which the most similar pair of documents or clusters is fused together to form a new cluster. For a collection of N documents, $N - 1$ fusions take place to result in a hierarchic classification in which small clusters of closely related documents are nested within larger and larger clusters of less related documents. A recent study has investigated four such hierarchic agglomerative clustering methods for automatic document classification [8,9]. Experiments were carried out to study the structures of the hierarchies produced by the different methods, the extent to which the methods distort the input similarity matrix during the generation of the classifications, and the retrieval effectiveness obtainable from searches of the clusters. The results suggested that the single linkage method, which has been used extensively in previous work on document clustering [5-7] and which has a well-developed theoretical basis, was not necessarily the most effective procedure of those tested.

This paper starts by continuing the comparison of these four hierarchic clustering methods. The results of the experiments suggest a simple but effective approach to nonhierarchic document clustering that is described in the second part of the report.

*To whom correspondence should be addressed.

Received May 28, 1985; revised June 17, 1985; accepted July 10, 1985

 CCC 0002-8231/86/010003-09$04.00

Data Sets and Evaluation Measures

Document Test Collections

The experiments used seven collections of documents, queries, and relevance judgements to ensure that the results were not unduly influenced by the characteristics of a particular data set. The collections were as follows.

(1) Keen. A set of 800 document titles, augmented by manually assigned indexing terms, and 63 queries on the subject of librarianship and information science.

(2) Cranfield. A set of 1400 documents and 225 queries on the subject of aerodynamics. These are characterized by lists of manually assigned index terms, whereas all of the following sets of documents and queries have been automatically indexed from natural language query statements and abstracts and/or titles.

(3) Evans. A set of 2542 document titles and 39 queries from the INSPEC data base that was used in an evaluation of search strategy variations in SDI profiles [11].

(4) Harding. A set of 2472 documents and 65 queries from the INSPEC data base that was used in an evaluation of automatic indexing techniques [12]. The documents used are a subset of those in the Evans collection, but with the titles augmented by abstracts to provide more exhaustive document characterizations and with a larger set of queries.

(5) LISA. A set of 6004 document titles and abstracts, the 1982 input to the Library and Information Science Abstracts data base, together with 35 queries. These were obtained from students and staff in this department, and the relevance judgements were obtained from manual searches of the printed version of the data base, supplemented in some cases by online or exhaustive manual searches [13].

(6) INSPEC. A set of 12,684 document titles and abstracts from the INSPEC data base, together with 77 queries collected at Cornell and Syracuse universities.

(7) UKCIS. A set of 27,361 document titles from the Chemical Abstracts Service data base, together with 182 queries collected by the United Kingdom Chemical Information Service in the early 1970s. This data set has been used extensively in previous document retrieval research but suffers from a lack of exhaustive relevance judgements.

In each case, the words in the document and query representatives were stemmed using a suffix-stripping algorithm after the elimination of common words on a stopword list. Duplicate stems were then eliminated, and the documents and queries were represented for search by lists of binary stem numbers.

The frequency characteristics of these collections are detailed in Table 1 where it will be seen that they span a wide range of types and data. Thus, there are long document descriptions with long (Harding) and short (Cranfield) queries as well as short documents with long (Evans) and short (UKCIS) queries. Moreover there are both broad queries (INSPEC and UKCIS) and sets of very specific queries with few relevant documents (Cranfield and LISA).

As well as detailing the frequency characteristics of each of the collections, Table 1 also contains *overlap* figures. These are derived from the cluster hypothesis test of van Rijsbergen and Sparck Jones [14], which states that similar documents tend to be relevant to the same requests. This hypothesis is both intuitively reasonable and easily tested for a document test collection by calculating all of the relevant-relevant (RR) and relevant-nonrelevant (RNR) interdocument similarity coefficients for some query: if the hypothesis is correct, it is to be expected that the RR coefficients will tend to be larger than the RNR coefficients. The results may be illustrated graphically by calculating the sets of RR and RNR coefficients for all of the queries in a collection and then plotting them as a pair of frequency distributions. The figures in Table 1 are the fractions of the two distributions that overlap each other; an example of such a plot, for the Cranfield data, is shown in Fig. 1, with the overlap area shaded. A collection with a low overlap value, such as Cranfield, will be one in which the relevant documents for the set of queries cluster strongly together and are well separated from the great bulk of nonrelevant material. Such collections are likely to be well suited to search strategies that are based upon the retrieval of clusters of documents, whereas collections with high overlap values, such as UKCIS or Evans, would seem to be inherently less well suited to such strategies.

Evaluation of Retrieval Effectiveness

The cosine coefficient [3] was used to determine the degree of similarity between a query and each of the clus-

TABLE 1. Characteristics of the document test collections.

	Keen	Cranfield	Evans	Harding	LISA	INSPEC	UKCIS
Number of documents	800	1400	2542	2472	6004	12684	27361
Number of queries	63	225	39	65	35	77	182
Number of terms per document	9.8	28.7	6.6	36.3	39.7	36.0	6.7
Number of terms per query	10.3	8.0	27.5	32.4	16.5	17.9	7.4
Number of relevant per query	14.9	7.2	23.1	22.6	10.8	33.0	58.9
Overlap	0.63	0.43	0.80	0.66	0.58	0.56	0.83

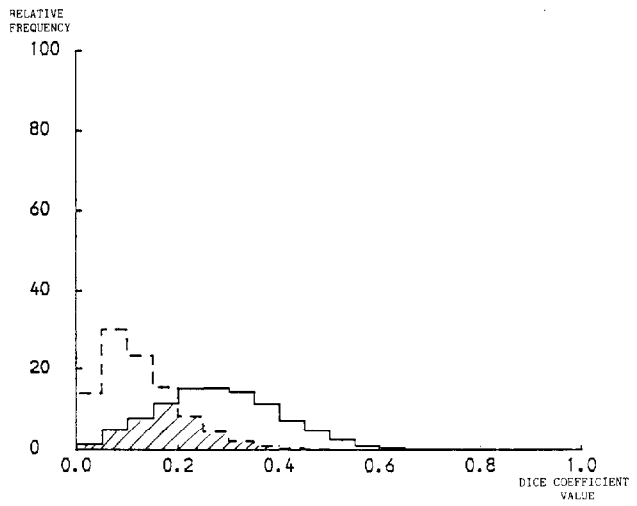

FIG. 1. Separation of RR and RNR distributions for the Cranfield test collection. The overlap area is shaded.

ters in a clustered search. For a cluster containing n_i occurrences of the ith term, the coefficient is defined as

$$\frac{\Sigma w_i n_i}{(\Sigma w_i^2 \Sigma n_i^2)^{1/2}}$$

where each of the query terms with a collection frequency of f_i in a collection of size N was assigned a weight, w_i, of

$$\log_e \frac{N}{f_i + 1},$$

and where the summation is over all of the terms in the indexing vocabulary. The clusters were ranked in descending order of similarity with each query for evaluation purposes.

The primary evaluation measure for the searches was the effectiveness measure, E [10]. For a search that retrieves a set of documents that give rise to recall and precision figures of R and P respectively, E is defined to be

$$1 - \frac{(1 + \beta^2)PR}{\beta^2 P + R}$$

where β is a user-defined parameter reflecting the relative importance attached to recall and to precision. A value for β of 0.5 (or 2.0) corresponds to attaching twice (or half) as much importance to precision as to recall, while a value of 1.0 corresponds to attaching equal importance to the two factors.

The success of the searches in providing relevance information was evaluated using two further measures. The first of these was the total number of relevant documents that were retrieved by the entire set of queries in some test collection, while the second of these was the number of queries in a collection for which the set of retrieved documents contained no relevant documents at all. These two measures will be denoted subsequently by T and Q, respectively. Thus, if R_i is the number of relevant documents retrieved in response to the ith query, T is the sum of all of the individual R_i values, while Q is the number of occasions for which R_i is zero. The reader should note that effective retrieval corresponds to low E or Q and high T values.

Two different searches of the same data set may be compared for significant differences by means of the sign test [10]. It is assumed that a large number of queries is available, so that the binomial distribution may be approximated by the normal distribution, and that the E values are available for the sets of documents retrieved by the two search strategies in response to each of these queries. Then, if C is the number of cases for which the two searches retrieve different numbers of relevant documents, and if c is the number of cases for which the first search strategy retrieves more relevant material than does the second strategy, then the test statistic that is evaluated is

$$\frac{c \pm 0.5 - 0.5C}{0.5\sqrt{C}},$$

where c is increased by 0.5 if it is less than $0.5C$ and decreased by the same amount if it is greater. The statistic follows the z distribution, and thus a calculated value greater than the critical value for z in a one-tailed test at some chosen level of significance, .05 in the work reported here, may be taken to imply that one search strategy gives significantly better retrieval than does the other.

Comparative Studies of Hierarchic Document Clustering Methods

The four hierarchic agglomerative clustering methods used here are all based upon the following simple algorithm:

(1) Calculate all of the interdocument similarity coefficients.
(2) Assign each document to its own cluster.
(3) Fuse the most similar pair of current clusters.
(4) Update the similarity matrix by deleting the rows and columns corresponding to the clusters that have been fused and calculating the entries in the row and column corresponding to the newly formed cluster.
(5) Return to step 3 if there is still more than one cluster.

This paper considers four such methods: single linkage, complete linkage, group average, and Ward's method. The methods differ in the updating mechanism used in step 4 above; a discussion of this is given by Lance and Williams [15], while Griffiths et al. [8] provide a summary of previous comparative studies of these methods.

Experiments were carried out using the small Keen, Cranfield, and Evans test collections with the classifications being generated by the CLUSTAN package [16]. Although widely available and flexible in operation [17], this package is restricted in the size of the data set that can be

processed. In practice, it was found that collections containing more than 800 documents could not be clustered, even when precomputed similarity matrices were used rather than the inefficient and time-consuming matrix generation routines in the package. Accordingly, only the Keen collection could be processed in toto; the Cranfield collection was hence split into two subsets, one of which contained the even-numbered documents and the other the odd; while a 1-in-4 systematic sample of the Evans documents was used for testing. The interdocument similarity measure that was used in the creation of the similarity matrices was the Dice coefficient.

Cluster Searches

The comparative studies in our previous paper [8] suggested that the single linkage classifications performed badly in optimal cluster searches for which full relevance data was available; however, the four methods gave more comparable levels of retrieval effectiveness in searches that retrieved a single cluster. These latter results are typified by the experimental runs reported in Table 2, which were obtained from searches of the bottom level clusters produced by the four methods. A *bottom level cluster* is the smallest cluster containing one of the documents in a collection, and thus corresponds to the cluster which that document joins when it first becomes connected into the cluster hierarchy [7]. Since it has been suggested that it is the small clusters that are important for good retrieval, a threshold cluster size of 40 documents was set so that only clusters smaller than this threshold size were included in the searches. The results are very similar to those reported previously [8] using a probabilistic model of cluster searching [7]; in particular, the group average method seems to perform consistently better than the other three methods in the $\beta = 2.0$ recall-oriented searches. This is confirmed by the sign test, since group average performs significantly better than all of the other methods with the two Cranfield subsets and better than single linkage when the Keen data are used.

Most experimental tests of cluster-based retrieval methods, such as those discussed above, have considered the retrieval of just a single cluster. However, this reflects a rather artificial retrieval environment since it would correspond to the retrieval of only two or three documents if a small bottom level cluster is identified as the best match for some query. In such cases, more than one cluster should be retrieved, and two sets of experiments were carried out to test the effect of retrieving additional documents. As before, the bottom level clusters were matched against each of the queries and ranked in descending order of the cosine coefficient. However, instead of retrieving

TABLE 2. Retrieval effectiveness of bottom level cluster searches.

	1 Cluster					5 Clusters					10 Documents				
	E Values					*E* Values					*E* Values				
Method	0.5	1.0	2.0	*T*	*Q*	0.5	1.0	2.0	*T*	*Q*	0.5	1.0	2.0	*T*	*Q*
							Keen								
Single linkage	0.77	0.83	0.85	77	26	0.79	0.81	0.81	142	17	0.81	0.83	0.82	138	19
Complete linkage	0.76	0.82	0.84	80	20	0.75	0.78	0.78	144	11	0.75	0.76	0.74	182	8
Group average	0.74	0.79	0.81	99	21	0.74	0.76	0.75	169	13	0.74	0.76	0.75	186	8
Ward's method	0.72	0.79	0.82	83	19	0.70	0.74	0.74	150	9	0.72	0.73	0.71	197	5
							Cranfield odd								
Single linkage	0.82	0.83	0.82	134	132	0.85	0.82	0.77	257	89	0.88	0.87	0.89	235	92
Complete linkage	0.79	0.80	0.80	129	118	0.80	0.77	0.71	255	77	0.84	0.80	0.72	297	63
Group average	0.78	0.78	0.76	186	109	0.81	0.77	0.70	302	72	0.84	0.80	0.72	302	69
Ward's method	0.79	0.81	0.81	120	122	0.80	0.77	0.72	236	78	0.85	0.81	0.73	288	66
							Cranfield even								
Single linkage	0.78	0.79	0.79	165	121	0.86	0.83	0.78	251	92	0.88	0.84	0.78	255	92
Complete linkage	0.78	0.80	0.80	140	120	0.81	0.78	0.73	257	83	0.84	0.80	0.72	323	65
Group average	0.75	0.76	0.75	211	107	0.82	0.78	0.72	318	73	0.83	0.79	0.71	334	71
Ward's method	0.77	0.79	0.79	145	119	0.81	0.78	0.73	259	85	0.84	0.80	0.72	325	65
							Evans								
Single linkage	0.83	0.87	0.88	30	22	0.85	0.84	0.82	54	12	0.85	0.84	0.81	55	12
Complete linkage	0.81	0.85	0.88	24	19	0.83	0.83	0.82	44	15	0.85	0.83	0.80	56	12
Group average	0.80	0.84	0.85	32	18	0.83	0.82	0.80	53	14	0.85	0.83	0.81	55	14
Ward's method	0.82	0.87	0.89	22	20	0.83	0.84	0.83	43	14	0.85	0.84	0.81	53	12

just the top-ranked cluster, either the 5 top-ranked clusters were retrieved or a sufficient number of clusters were retrieved to give a total of 10 distinct documents. If a greater number was obtained in the latter case, sufficient documents were randomly selected from the last of the retrieved clusters to ensure that all of the searches resulted in exactly the same fixed number of documents. It is felt that such searches provide a more realistic comparison of the merits of the different clustering methods than do experiments that involve the retrieval of just a single cluster. The results of the experiments using the top 5 clusters and top 10 documents are included in the right-hand portions of Table 2.

Few statistically significant differences in performance are evident in the case of the Evans collection; this may be due to the high overlap figure since, if there is little separation between the relevant and nonrelevant documents, there is unlikely to be large differences in retrieval effectiveness when the file is clustered for search in different ways. Both halves of the Cranfield collection show single linkage to be significantly worse than the other three methods for the retrieval of either 5 clusters or 10 documents. This is also the case for the retrieval of 10 documents from the Keen data; and for this test set, Ward's method gives significantly better results than does complete linkage when 5 clusters are retrieved. Griffiths and Willett [9] report experiments in which the use of the cosine coefficient was replaced by the use of a probabilistic cluster search. The results were very similar to those reported here, the only noticeable difference was that Ward's method was often significantly better than all of the other three methods and not just single linkage.

In summary, the cluster searches carried out here and previously would suggest that the single linkage method results in searches that are sometimes inferior to those obtained from the use of the other three methods; of these, Ward's method may give the best results if more than a single cluster is to be retrieved. However, it should be emphasized that the experiments have involved only a single basic search mechanism, and that different results might be obtained if, for example, top-down or bottom-up searches [6] of the full cluster hierarchies were to be undertaken.

Cluster Membership

A second set of experiments involved an investigation of the size and constitution of the bottom level clusters produced by the four methods.

The distribution of sizes in the bottom level clusters is shown in Table 3. Complete linkage, group average, and Ward's method show a similar distribution, with about three-quarters of the clusters containing just a pair of documents and with very few clusters containing more than 10 documents. Single linkage shows a quite different pattern of behavior, with a much less skewed distribution of cluster sizes and with very many large clusters. Thus, over 37% of the bottom level Keen clusters contain more than 400 documents, and over 10% contain more than 700; similar behavior is observed with the other test sets. Such a distribution of cluster sizes is a natural reflection of the highly unstructured character of single linkage hierarchies that has been noted in previous work [8,18].

TABLE 3. Distribution of sizes of bottom level clusters.

Method	Cluster Size					
	2	3	4	5-20	21-40	>40
	Keen					
Single linkage	234	74	30	59	8	395
Complete linkage	598	141	34	25	0	2
Group average	556	125	48	67	2	2
Ward's method	634	130	30	6	0	0
	Cranfield odd					
Single linkage	230	56	26	35	18	335
Complete linkage	520	126	30	23	0	1
Group average	468	125	43	60	2	2
Ward's method	546	126	18	10	0	0
	Cranfield even					
Single linkage	252	50	21	48	10	319
Complete linkage	540	115	26	17	0	2
Group average	478	121	41	55	3	2
Ward's method	550	121	21	8	0	0
	Evans					
Single linkage	184	58	25	51	3	314
Complete linkage	448	94	35	47	0	11
Group average	416	101	42	68	4	4
Ward's method	494	106	21	14	0	0

The figures in Table 3 raise some questions about the retrieval results in Table 2 since it may appear that the experiments had been biased against the single linkage method and toward clustering methods that tend to produce large numbers of small clusters. This is because the use of a threshold cluster size of 40 documents excludes considerable portions of the single linkage classifications from consideration during a search, whereas the great bulk of the bottom level clusters for the other three methods are smaller than this threshold size. To test whether the single linkage results were being affected, searches were carried out on the Keen data in which the threshold bottom level cluster size was progressively increased so that a greater and greater fraction of the file was available for search. The $\beta = 0.5$ and $\beta = 1.0$ searches showed a marked and progressive decrease in effectiveness as the threshold size was increased, although the $\beta = 2.0$, recall-oriented searches were less affected by the increase in the mean cluster size. The T and Q figures revealed an increasing amount of relevance information, but this was obtained only at the expense of a quite drastic increase in the numbers of documents retrieved as the mean cluster size

grew. It would thus seem that the chosen methodology does not seriously disadvantage the single linkage method.

Two further points are of importance. First, it is intuitively reasonable that the larger a cluster becomes, the less accurately the representative describes the documents that are contained within that cluster, and it is thus to be expected that small clusters will give better search performance than do larger ones. This would certainly appear to be the case in the experiments reported here, while Croft [7] has reported results which confirm this expectation in an extended series of cluster based retrieval experiments using a single linkage classification of the full Cranfield test collection. The second point that needs to be made is a consideration of how an operational retrieval system based on document clusters might function. As noted in the first section of this paper, cluster searching has been advocated as a means of obtaining a few relevant documents that may then be used as the basis for a relevance feedback search. The feedback is based on user judgements of the relevance of the few documents retrieved in an initial search; thus, a cluster search that retrieved a very large cluster will require either very many relevance judgements from the user, which he or she may well not wish to provide, or a means for the selection of some small number of documents from the cluster. This latter approach may be accomplished by a variety of means—such as the matching of the individual documents in a cluster against the query, or the selection of those documents most similar to the representative—but this is at variance with the aim of retrieving document clusters in their entirety. Taking these two points together, it would seem that clustering methods that result in small numbers of large clusters are inherently less suitable for cluster based retrieval than are methods that result in large numbers of small clusters.

Luckhurst [19] describes additional experiments in which she measured the degree to which the same bottom level clusters were identified using different clustering criteria. A large degree of overlap was found among the complete linkage, group average, and Ward clusters that share many small clusters in common; for each of the test collections studied, about 75% of the bottom level clusters identified by these three methods were the same, whereas only about 40% of the single linkage clusters were identical with those produced by any one of the other three methods. This finding is again in line with other findings that single linkage gives classifications that yield rather different search results from the other three types of cluster.

Use of Nearest Neighbor Clusters

It must be emphasized that the results presented in the previous section have been obtained using very small sets of documents, and it is clear that the experiments should be repeated using significantly larger data sets. Until such tests have been completed, it may be noted that the results to date suggest that the best searches are obtained from the use of clusters containing only small numbers of documents. The smallest such clusters will contain just a document and its nearest neighbor, i.e., that document with which a specified document has its greatest similarity, and this section investigates the use of such *nearest neighbor clusters* (hereafter NNCs) for document retrieval. As before, the Dice coefficient was used for the determination of all of the interdocument similarity coefficients.

The organization of a file of documents on the basis of the NNCs represents an overlapping classification, since a given document may occur in more than one cluster. It should be noted that if a pair of documents are reciprocal nearest neighbors, i.e., if document j is the nearest neighbor of document i, and i is the nearest neighbor of j, the NNCs for i and j will be identical and only one cluster need be stored for search. Such occurrences will mean that, in general, less than N clusters need to be inspected in an NNC search of a file containing N documents.

The use of nearest neighbors has figured prominently in the general clustering literature [20-22] as well as in the specific context of document classification. Thus Goffman [23] and Mansur [24] have discussed retrieval methods based on chains of nearest neighbors, Willett [25] has considered using sets of nearest neighbors for generating single linkage clusters, while Croft [26] and Croft et al. [27] have described a network organization for information retrieval in which both documents and terms are linked to their nearest neighbors. However, these reports have not involved detailed retrieval experiments using document collections of realistic size; such tests are described below.

Effectiveness of Retrieval

A limitation of the work described in the previous section is that it considers only the comparison of one cluster search with another, without considering the retrieval effectiveness obtainable from conventional best match searching. Accordingly, the NNC searches are compared with *full searches* in which the queries are matched against each of the documents in the file. The full search is based on the collection frequency weights detailed above, with the similarity between a document and a query being calculated by the sum of the weights for the matching terms. The documents were ranked in descending order of similarity with each of the queries, and a threshold of 10 or 20 documents applied to the ranking to obtain a set of documents for the measurement of retrieval effectiveness. An entirely comparable procedure was used with the NNCs, these being ranked in descending order of the cosine match used in the previous section; sufficient clusters were then retrieved to obtain either 10 or 20 documents as required.

The effectiveness of the two types of search are detailed in Table 4. With one or two exceptions, the overwhelming impression is one of little or no difference between the two types of search strategy, with both giving similar levels of

TABLE 4. Retrieval effectiveness of full and NNC searches using a cutoff of 10 or 20 documents.

	Full Search					NNC Search				
	E Values					*E* Values				
	0.5	1.0	2.0	*T*	*Q*	0.5	1.0	2.0	*T*	*Q*
	10 documents									
Keen	0.73	0.74	0.72	186	4	0.72	0.73	0.71	202	6
Cranfield	0.80	0.78	0.73	433	52	0.75	0.73	0.68	533	35
Evans	0.78	0.83	0.85	113	3	0.80	0.84	0.85	103	4
Harding	0.83	0.86	0.88	155	19	0.83	0.87	0.89	149	24
LISA	0.80	0.80	0.78	74	9	0.79	0.80	0.77	78	7
INSPEC	0.80	0.85	0.87	233	10	0.83	0.87	0.89	203	11
UKCIS	0.89	0.91	0.92	340	75	0.90	0.93	0.94	316	77
	20 documents									
Keen	0.77	0.75	0.69	280	2	0.77	0.75	0.69	289	3
Cranfield	0.84	0.80	0.72	630	29	0.82	0.77	0.67	732	21
Evans	0.80	0.81	0.81	170	1	0.80	0.81	0.81	164	2
Harding	0.84	0.85	0.85	236	15	0.84	0.86	0.86	227	20
LISA	0.83	0.81	0.76	112	5	0.83	0.81	0.75	113	5
INSPEC	0.80	0.82	0.83	370	3	0.83	0.85	0.86	325	5
UKCIS	0.88	0.90	0.90	564	60	0.90	0.91	0.91	513	54

performance. This impression is confirmed, in general by the use of the sign test, since differences at the .05 level of statistical significance are observed only for the Cranfield and threshold-20 INSPEC searches. The Cranfield NNC results are significantly better than the full searches, even at the .0001 level of significance, and it is clear that the low overlap value for this collection is reflected in quite excellent cluster searches. It may be noted in passing that the difference between the two types of search is even more marked if a threshold of 5 documents is used, with the NNC search here retrieving some 35% more relevant documents than the full search and giving a $\beta = .05E$ value as low as 0.71. In the case of the INSPEC data, the threshold-20 NNC searches are significantly inferior to the full search at the .005 level of significance.

Early work on document clustering [28] found that cluster searches were markedly less effective than full searches. While more recent studies [3,29] have suggested that the two types of search may be rather less disparate in performance, the results obtained here do provide some form of justification for the use of clusters for organizing a document collection. Moreover, the results are acceptable even with the Evans and UKCIS collections where the overlap figures suggest that the data may not be amenable to a clustered organization. The most interesting results are those for the Cranfield collection since the NNC searches are far superior not only to the full search here but also to all of the strategies used by Croft and Harper in their studies of probabilistic searching in the absence of relevance information [3]. The NNC results are also noticeably better than those reported for bottom level cluster searches of a single linkage classification of this data set [7].

Combining the Two Types of Search

It has become increasingly clear that different search mechanisms result in the retrieval of quite different sets of documents [30], and it has accordingly been suggested that future document retrieval systems should incorporate a range of search strategies that can be selected, either by the system or by a user, as appropriate to the needs of a particular query [26]. That such a strategy can indeed increase the effectiveness of retrieval is shown by the "optimal" results listed in Table 5 where the full and NNC searches have been compared and the evaluation measures calculated using that type of search that gives the better result for each of the queries. The results are, of course, very much better than those listed in Table 4, not only in terms of the total numbers of relevant documents but also in terms of the queries retrieving relevant material since there are many cases where the full search retrieves at least one relevant document whereas the NNC search does not, and vice versa. Only in the case of the Cranfield data is there little improvement when the optimal results are compared with the individual types of search, this exceptional behavior arising from the fact that the NNC searches for this collection are so good that little benefit accrues from providing the full search as an alternative retrieval mechanism.

An analysis of the output from the two types of search shows that, although the total numbers of relevant document retrieved by the two types of search are very similar, the two sets of output often have relatively few relevant documents in common. For example, the full and NNC threshold-10 searches of the INSPEC data retrieved 233 and 203 documents, respectively, but only 90 of these were

TABLE 5. Retrieval effectiveness of combined and optimal full and NNC searches using a cutoff of 10 or 20 documents.

	Optimal Search					Combined Searth				
	E Values					*E* Values				
	0.5	1.0	2.0	*T*	*Q*	0.5	1.0	2.0	*T*	*Q*
	10 documents									
Keen	0.69	0.70	0.67	224	3	0.72	0.73	0.71	198	4
Cranfield	0.74	0.71	0.65	568	25	0.76	0.74	0.69	515	35
Evans	0.75	0.80	0.82	128	1	0.78	0.82	0.84	110	4
Harding	0.80	0.85	0.87	179	17	0.82	0.86	0.88	161	20
LISA	0.77	0.77	0.73	87	5	0.79	0.79	0.76	81	7
INSPEC	0.77	0.83	0.86	269	3	0.80	0.85	0.88	235	8
UKCIS	0.87	0.90	0.91	421	55	0.89	0.92	0.92	349	67
	20 documents									
Keen	0.74	0.72	0.65	324	2	0.77	0.75	0.69	291	2
Cranfield	0.81	0.76	0.65	774	13	0.82	0.78	0.68	709	21
Evans	0.77	0.78	0.78	195	1	0.79	0.80	0.80	174	1
Harding	0.82	0.83	0.83	264	14	0.83	0.84	0.85	246	17
LISA	0.81	0.79	0.72	126	3	0.83	0.81	0.75	114	5
INSPEC	0.78	0.80	0.81	413	1	0.80	0.83	0.83	370	3
UKCIS	0.86	0.88	0.88	672	38	0.89	0.90	0.90	556	49

common to both types of search. This suggests that an improved level of performance might be achieved from a retrieval strategy that encompassed both types of search so that some of the documents presented to a user had come from the full search and some of them from the NNC search. In the absence of any obvious rule as to what proportion of the documents should come from each type of search, and as to how this proportion should vary from one query to another, sets of 10 (or 20) documents were obtained for performance evaluation by merging the top 5 (or 10) documents from the full search ranking with the top 5 (or 10) documents from the NNC ranking (after the elimination of any duplicates). The results of these "combined" searches are listed in Table 5 and may be compared with the corresponding figures from Table 4. No large differences in performance can be seen with the exception of the UKCIS collection, where noticeably fewer queries retrieved no relevant material in the combined searches, and of the Cranfield data, where the NNC results are so good that the inclusion of material from a full search proves to be deleterious.

An alternative, and more sophisticated, means of combining two, or more, types of search is suggested in a recent paper by Croft and Thompson [31], who describe an adaptive mechanism that tries to learn which retrieval strategy is most appropriate for a given query. Unfortunately, the tests showed that, although the approach had some merit, it was not possible to obtain results that were superior to those obtainable from the consistent use of just a single strategy.

Implementation Details

The problem of identifying the document(s) most similar to some query, the nearest neighbor problem, has been intensively studied over the last few years, and several inverted file algorithms have been described that may be used for this purpose. The algorithm used here was that described by Noreault et al. [32]. This involves the addition of the inverted file lists corresponding to the terms in the query to yield a vector, the ith element of which contains the sum of the weights of the terms common to the query and to the ith document. The largest such element then specifies the nearest neighbor for that query. An example of an operational retrieval sysem based on this algorithm is given by Brzozowski [33], and it may be also used for the generation, search, and updating of a file of NNCs.

The clusters may be generated by using the algorithm to identify the nearest neighbor of each of the documents in a collection [34]. In such a case, the algorithm will have a running time of order $O(N^2)$, but the constant of proportionality is sufficiently small to make the algorithm practicable for files of nontrivial size. Our experiments used an elderly ICL 1906S computer with the programming in Algol 68, not a particularly efficient language. The identification of the 20 nearest neighbors, rather than just the single nearest neighbor as used here, for each of the documents in the SMART and UKCIS collections required about 5-1/2 and 3 hours of CPU time, respectively. This would suggest that the NNCs for files of up to 100,000 documents should be obtainable using a modern mainframe and assembly-level coding of the algorithm.

A further advantage of using an inverted file to support the searching of the NNCs is that, unlike most document clustering schemes that have been suggested, there are no overheads associated with the storage of cluster centroids, or representatives, since the requisite information is available from the inverted file. The only storage requirements additional to those of a conventional full search is an N-component array, the kth component of which contains the identifier for the nearest neighbor of the kth document, and the Σn_i^2 term in the denominator of the cosine coefficient for the kth NNC.

With one or two slight modifications, the algorithm may also be used to search the file of NNCs. This search may be carried out at the same time as a full search so that, once the set of NNCs has been generated, NNC searching involves little more processing than does a conventional full search of a document collection. Updating the lists of nearest neighbors as a new document, k, is added to the collection involves using the algorithm to calculate the similarity between k and each of the current members of the collection: the largest similarity corresponds to the nearest neighbor for k, while k becomes the nearest neighbor for some document l if the similarity between l and k is greater than that between l and its current nearest neighbor.

Very similar conclusions have been reached by Croft et al. [27] in a detailed study of the generation, search, and updating requirements of a network file structure that contains both documents and terms, rather than just documents as in the work reported here. These authors also report a simulation study of the numbers of disk accesses that are required for the searching of the network, and their results would be applicable, in large part, to a retrieval system that was based upon NNCs.

Conclusions

The first part of this article describes a comparison of four different hierarchic agglomerative clustering methods. These experiments show that the best results, in terms of the effectiveness of retrieval, are given by clustering methods that result in large numbers of small clusters. Since many of these clusters will contain just a document and its nearest neighbor, the findings would suggest the use of a file organization based upon clusters that contain a document and its nearest neighbor, an NNC. Such files may be generated, searched, and updated at a relatively low computational cost.

Searches of the sets of NNCs for seven document test collections were shown to give a level of retrieval effectiveness little different from that obtainable in conventional full searches. However, an analysis of the actual docu-

ments that were retrieved by the two types of search shows that they identify rather different sets of relevant documents. This would suggest that a retrieval system should contain both sorts of retrieval mechanism if it is to provide an effective response to as wide a range of queries as possible. Such a system would use one of the search types as the basic strategy, but could then switch to the other if the initial set of documents proved to be unsatisfactory.

Attempts were made in the experiments to combine the outputs from the two types of search into a single set of documents, but this was found to be less successful than a strategy that carried out both types of search and then selected the more effective one to provide the output. Such a retrospective approach can, of course, be used only in an experimental environment where full relevance information is available, and an operational retrieval system would need to have some selection mechanism that would allow it to select which strategy should be applied to a particular query. Studies are now in progress to identify appropriate selection methods for this purpose.

Acknowledgments

Thanks are due to Dr. K. Sparck Jones (Cambridge University), Prof. G. Salton (Cornell University), Mr. L. Evans, Mr. P. Harding, and Mr. J. Pache (INSPEC), and Mr. N. Moore (Library Association) for providing the document test collections used in this study. Funding was provided by the British Library Research and Development Department under grant number SI/G/564 and by the award of a Department of Education and Science Advanced Course Studentship to H. C. L.

References

1. Robertson, S. E.; Sparck Jones, K. "Relevance weighting of search terms." *Journal of the American Society for Information Science*. 27:129-146; 1976.
2. Harper, D. J.; van Rijsbergen, C. J. "An evaluation of feedback in document retrieval using co-occurrence data." *Journal of Documentation*. 34:189-216; 1978.
3. Salton, G.; McGill, M. J. *Introduction to Modern Information Retrieval*. New York, NY: McGraw-Hill; 1983.
4. Croft, W. B.; Harper, D. J. "Using probabilistic models of document retrieval without relevance information." *Journal of Documentation*. 35:285-295; 1979.
5. Jardine, N.; van Rijsbergen, C. J. "The use of hierarchical clustering in information retrieval." *Information Storage and Retrieval*. 7:217-240; 1971.
6. van Rijsbergen, C. J.; Croft, W. B. "Document clustering: an evaluation of some experiments with the Cranfield 1400 collection." *Information Processing and Management*. 11:171-182; 1975.
7. Croft, W. B. "A model of cluster searching using classification." *Information Systems*. 5:189-195; 1980.
8. Griffiths, A.; Robinson, L. A.; Willett, P. "Hierarchic agglomerative clustering methods for automatic document classification." *Journal of Documentation*. 40:175-205; 1984.
9. Griffiths, A.; Willett, P. *Evaluation of Clustering Methods for Automatic Document Classification*. London, UK: British Library Research and Development Department; 1984.
10. van Rijsbergen, C. J. *Information Retrieval*. London, UK: Butterworth; 1979.
11. Evans, L. *Search Strategy Variations in SDI Profiles*. London, UK: INSPEC; 1975.
12. Harding, P. *Automatic Indexing and Classification for Mechanised Information Retrieval*. London, UK: INSPEC; 1982.
13. Davies, A. *A Document Test Collection for Use in Information Retrieval Research*. M.S. dissertation, University of Sheffield, UK; 1983.
14. van Rijsbergen, C. J.; Sparck Jones, K. "A test for the separation of relevant and non-relevant documents in experimental retrieval collections." *Journal of Documentation*. 29:251-257; 1973.
15. Lance, G. N.; Williams, W. T. "A general theory of classificatory sorting strategies. I. Hierarchical systems." *Computer Journal*. 9:373-380; 1967.
16. Wishart, D. *CLUSTAN 1C User Manual*. Edinburgh, UK: Edinburgh University Program Library Unit; 1978.
17. Aldenderfer, M. S. *A Consumer Report on Cluster Analysis Software*. University Park, PA: Pennsylvania State University; 1977.
18. Murtagh, F. "Structure of hierarchic clusterings: implications for information retrieval and multivariate data analysis." *Information Processing and Management*. 20:611-617; 1984.
19. Luckhurst, H. C. *A Comparison of Four Hierarchical Clustering Methods for Document Retrieval*. M.S. dissertation, University of Sheffield, UK; 1984.
20. Jarvis, R. A.; Patrick, E. A. "Clustering using a similarity measure based on shared nearest neighbours." *IEEE Transactions on Computers*. C-22:1025-1034; 1973.
21. Gowda, K. C.; Krishna, G. "Agglomerative clustering using the concept of mutual nearest neighbourhood." *Pattern Recognition*. 10:105-112; 1978.
22. Mizoguchi, R.; Shimura, M. "A nonparametric algorithm for detecting clusters using hierarchical structure." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. PAMI-2:292-300; 1980.
23. Goffman, W. "An indirect method of information retrieval." *Information Storage and Retrieval*. 4:363-373; 1969.
24. Mansur, O. "An associative search strategy for information retrieval." *Information Processing and Management*. 16:129-137; 1980.
25. Willett, P. "A note on the use of nearest neighbours for implementing single linkage document classifications." *Journal of the American Society for Information Science*. 35:149-152; 1984.
26. Croft, W. B. "Incorporating different search models into one document retrieval system." *ACM SIGIR Forum*. 16:40-45; 1981.
27. Croft, W. B.; Wolf, R.; Thompson, R. "A network organization used for document retrieval." *ACM SIGIR Forum*. 17:178-188; 1983.
28. Salton, G. *The SMART Retrieval System*. Englewood Cliffs, NJ: Prentice-Hall; 1971.
29. Willett, P. "Document clustering using an inverted file approach." *Journal of Information Science*. 2:223-231; 1980.
30. Katzer, J.; McGill, M. J.; Tessier, J. A.; Frakes, W.; DasGupta, P. "A study of the overlap among document representatives." *Information Technology: Research and Development*. 1:261-274; 1982.
31. Croft, W. B.; Thompson; R. H. "The use of adaptive mechanisms for selection of search strategies in document retrieval systems." In: van Rijsbergen, C. J. (Ed.) *Research and Development in Information Retrieval*. Cambridge, UK: Cambridge University Press; 1984.
32. Noreault, T.; Koll, M.; McGill, M. J. "Automatic ranked output from Boolean searches in SIRE." *Journal of the American Society for Information Science*. 28:333-339; 1977.
33. Brzozowski, J. P. "MASQUERADE: searching the full texts of abstracts using automatic indexing." *Journal of Information Science*. 6:67-73; 1983.
34. Willett, P. "A fast procedure for the calculation of similarity coefficients in automatic classification." *Information Processing and Management*. 17:53-60; 1981.