VII. A Modified Two-Level Search Algorithm Using Request Clustering

V. R. Lesser

1. Introduction

In the past few years, prototype time sharing computer systems have been developed which have made it possible to obtain access to computers by remote console. In the context of an information retrieval system, this development is likely to affect the systems operations: from a batch type processing of queries to single query processing introduced into the system via remote console. Of still greater importance is the fact that this change makes possible the use of an information retrieval system by a large and diverse user population. Because of these new developments in computer organization, a considerable degree of emphasis has been placed on procedures for using a system of man-machine interaction to improve the retrieval of relevant documents in answer to search requests from a population of users [4,5,6]. Such a change of procedure necessitates a redesign of the techniques of document retrieval to make them adaptable to a single query processing environment.

In a batch processing organization, it is not unreasonable to wait until a large set of queries accumulates, and thereafter to search the whole document collection in one pass to identify documents which are highly correlated with the batch of queries. In a real time system, on the other hand, queries cannot be batched; as a result a search of the whole document collection for each query becomes very uneconomical, and the

full search technique must be discarded. Instead, a multi-level search algorithm based on a partial scan of the document collection can be used. Such a scheme can be based on a clustering algorithm which uses the information content of the documents to partition the document collection into subsets of related documents. The following procedure can then be used to perform a multi-level search based on the previously identified document clusters:

1) the procedure finds those subsets of documents whose representative centroid vectors* are significantly correlated with the given query;

2) the query is then matched against each document contained in the subsets of documents found in step 1).

The basic assumption in the multi-level search is that for each new query introduced into the system, the documents which are relevant to this query are contained in only a few of the document clusters. Further, the centroid vectors of these particular clusters will correlate more highly with the query than the centroid vectors of the document clusters which do not contain any documents relevant to the query. Therefore, the efficiency of the multi-level search is dependent on the number of clusters which contain relevant documents, and on the correlation of the query with the centroid vectors of relevant document clusters.

---

* (let each document be represented by an n- dimensional index vector; consider a set D of document index vectors; the centroid vector $\underline{c}$ for the set D is defined as:

$$c = \frac{1}{n} \sum_{i=1}^{n} \frac{\overline{di}}{|di|}$$

where $D = \{\overline{d}_1, \overline{d}_2, \ldots\ldots, \overline{d}_n\}$

It is felt that the partitioning of the document collection by grouping documents containing similar information identifiers does not always maximize the efficiency of the multi-level search. This technique of partitioning is effective when the set of queries introduced into the system can be divided into groups of queries which roughly correspond in information content to the subsets of documents previously created by the clustering algorithm. If this is not the case, the set of relevant documents for a query will be spread over many document subsets, and the multi-level search will not prove effective. In practice, it is believed that the distribution of the information content of the queries may often differ significantly from that of the document collection. Furthermore, if this contention is correct, a more efficient classification scheme can be constructed by considering the information content of queries previously introduced into the system.

In the next few paragraphs, new techniques are described for partitioning the document collection, and for carrying out the multi-level search, in accordance with the query set previously introduced into the system, as well as a possible modification of this technique of partitioning based on relevance judgments provided by the user.

2. A Modified Clustering Algorithm and a Corresponding Two-Level Search Strategy

It is desired to construct clusters of documents as a function of both the collection of documents, and also the collection of previous queries introduced into the system. The procedure for clustering is divided into three stages:

1) the collection of previous queries introduced into the system is partitioned into subsets of queries using a standard clustering algorithm;[1,6]

2) an associated subset of documents is formed for each subset of queries constructed in step 1); the associated subset of documents consists of all documents which are highly correlated with at least one query contained in the subset of queries;

3) all documents which are not associated with any query cluster by step 2) are divided into subsets using a standard clustering algorithm.

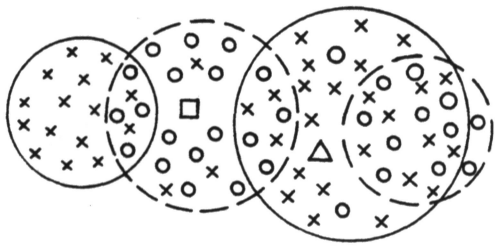
The multi-level search previously described is then modified to take into account this new request clustering procedure. The new modified two-level search algorithm uses the following procedure: the new query is correlated against the centroid vectors of the cluster subsets of previous queries; if the new query correlates highly with at least one of the query centroid vectors, the query is matched against each document contained in the associated subsets of documents corresponding to each highly correlated query centroid vector; otherwise, the new query is matched against the centroid vectors of the subsets of non-associated documents constructed in step 3); for those subsets whose centroid vector correlates highly with the query, the query is matched against every document contained in the subset.

This new clustering algorithm can be further modified by incorporating user relevance judgments for each previous query introduced into the system. In step 2), instead of associating all those documents which were identified by their high correlation, it is possible to associate only those documents considered relevant to the query by the user.

3. Advantages of the Query Clustering System

The modified two-level search algorithm should in practice be more efficient than the normal multi-level search, since fewer subsets of documents will have to be compared completely with the requests in the former. This may be expected to be true because a new query will not be matched against the centroid vectors of information dense subsets of documents, but against the centroid vectors of subsets of previous queries to the system. The matching of the new query against the centroid vectors of previous queries should cause a natural association of this new query with one particular subset of queries, and therefore fewer documents will have to be compared to satisfy the user's request. An illustration of this situation is shown in Figure 1.

Obviously, in an information retrieval system where the document collection because of its size must be stored in external memory (e.g. disk, drum, data cell), the number of categories which need to be completely searched becomes a critical time factor. Each time a new category is searched, the documents which are associated with this category must be obtained from external storage. This data handling operation requires a considerable amount of time, so that its minimization is an important factor in speeding up the retrieval process.



x represents a document
o represents a query
___ represents a document cluster
--- represents a query cluster and the subset of documents which is associated with the query cluster

Advantages of Query Clustering

Figure 1

Consider as an example, a new query (□) introduced into the system, and assume that the new query lies between two document clusters. In the normal two level search scheme both document clusters will therefore have to be searched completely. However since the given query lies in a cluster of previous queries the centroid vector of only one query cluster will correlate highly with it. Therefore in the modified two-level search scheme, only the subset of documents which is associated with the one highly correlated query cluster needs to be searched completely. The question now arises as to why a query (△) which lies between two query clusters could not occur with the same frequency as the query (□) used in the example. Two basic assumptions underlie the modified two level search:

1) The set of previous query vectors to the system form dense subsets in the n-dimensional space so that the set of previous queries can be clustered into subsets based on the information content of each query;

2) A new query is on a statistical basis likely to be similar to a subset of previous queries so that the new query can be assigned to a subset of previous queries created by the clustering algorithm.

If these two assumptions are true, it is much less likely for a new query to lie between two query clusters than between two document clusters.

4. Design of an Experiment to Compare the Modified with the Normal Two-Level Search Scheme

A) Problem Areas

In structuring this experiment, the following questions must be answered:

1) What criteria can be used to judge the relative merits of each procedure ?

2) How should these alternative search procedures be implemented, and what parameters must be adjusted in using these procedures ?

3) What type of document and query collection will serve as an adequate data base in order to obtain valid conclusions ?

B) Tests to Compare the Effectiveness of Each Search Procedure

The main criterion for effectiveness is based on the number of documents which must be scanned in each procedure in order to obtain most of the relevant documents for each query. In a practical implementation of a normal two level search scheme, the number of subsets completely searched will be either a fixed number for all queries, or will depend on the correlations with the query of the centroid vectors of the document subsets. Neither of these procedures for determining the number of categories to be searched completely can be used to compare the effectiveness of the modified two-level search scheme with the normal two-level search, since neither the number of clusters nor the size of each cluster is the same for both search schemes. These differences in the number of clusters, and the size of each cluster make it impossible to use the same parameters for determining the number of subsets of documents to be completely searched for each of the search procedures. Further,

these parameters are arbritary so that in order to validly compare alternative search procedures, the parameters would have to be adjusted to maximize the effectiveness of each search procedure. Therefore, a different algorithm which is not a function of the number of clusters nor the size of a cluster is used to calculate the number of documents to be completely searched.

In order to generate the criterion for search effectiveness, the normal procedure for querying a document collection is altered: instead of considering a user request consisting of only a query together with a cut-off value for the correlation coefficient (only documents which correlate above the cut-off value are retrieved for each query), an additional parameter is included. This parameter specifies the number of documents to be retrieved. In this modified querying system, each search procedure is altered so that when the specified number of documents are retrieved, the search procedure terminates. This modification permits the comparison of the minimum number of documents each search procedure must scan in order to satisfy the modified user request. There also must be available some measure of the extent of relevance of the documents retrieved by the alternative search procedures in relation to the documents retrieved by a full search of the document collection.

Rocchio [6] in comparing the effectiveness of a two-level search algorithm based on his clustering algorithm against the effectiveness of a full search of the document collection uses the following criteria:

1) the "consistency of retrieval with respect to all documents," i.e. the extent to which the reduced search leads to the retrieval of the same documents as the full search;

2) the "consistency of retrieval with respect to relevant documents," i.e. the extent to which the retrieval of the relevant documents is altered by the reduced search.

The above criteria are based on the amount of information lost when the documents are retrieved by a partial search of the document collection instead of by a full search. It is believed that in conjunction the two criteria for effectiveness provide adequate data for an appraisal of the modified two-level search scheme compared with the normal two-level search scheme.

In the modified querying system proposed for testing, Rocchio's two criteria take the following form:

1) the overlap percentage between the retrieved set of documents obtained by the partial search with the first $\underline{n}^*$ documents retrieved by the full search;

2) the normal recall or the percentage of relevant documents retrieved by the partial search to the number of relevant documents contained in the first $\underline{n}^*$ documents retrieved by the full search.

C) Implementation of the Normal and Modified Two-Level Search Schemes

Each search procedure relies heavily on the particular clustering algorithm used, and the parameters used by the cluster algorithm to determine how the document collection is to be partitioned. It was decided, based on a search of the literature, that Rocchio's clustering algorithm [6] would be the most suitable. The parameters that are used

---

* n = the number of documents to be retrieved originally specified by the user for the partial search of the document collection.

by this cluster algorithm are:

1) the number of partitions (categories);
2) the minimum and maximum size of a category;
3) the parameters which define an acceptable category (the density test).

Normal Two-Level Search:

The data required to carry out this process are completely generated by the application of the clustering algorithm to the document collection. Certain documents called "loose" will not be classified into any category. In order to have the same documents included in the set of document clusters constructed for each search procedure, a loose document is associated with that category whose centroid vector exhibits the maximum correlation with the given loose document. The parameters that can be varied in the construction of the document clusters for the normal two-level search are:

1) the number of document clusters;
2) the size of a cluster;
3) the parameters for the density test.

Modified Two-Level Search:

The implementation proceeds in three steps:

1) the collection of previous queries introduced into the system is partitioned using the standard clustering algorithm, and all loose queries are eliminated since they are statistically of no consequence; the composition of the query clusters can be varied in the following manner:
    a) the number of query clusters;

b) the size of each query cluster;

c) the criterion for an acceptable query cluster.

2) the document collection is partitioned into a set of associated and non-associated documents based on the query clusters in step 1); the formation of subsets of associated documents for each given query cluster can take place in one of three ways:

a) the associated subset of documents for the given query cluster is formed by associating documents which correlate highly with a query contained in the given query cluster;

b) the associated subset of documents for the given query cluster is formed by associating documents which correlate highly with the centroid vector of the given query cluster;

c) the associated subset of documents for the given query cluster is formed by associating documents which are judged by a user relevant to his query contained in the given query cluster.

The size of the subsets of associated documents depends on what is meant by "highly correlated"; the size can be determined in one of two ways:

a) the size depends on the number of queries contained in the given query cluster in the sense that the greater the number of queries contained in a query cluster the greater the number of documents that are associated with a query cluster (this method of determining the size of the associated categories is rationalized by the expectation that certain areas of information will more often contain relevant documents for the query, so that these information areas should be larger);

b) the size depends on the density of the documents which surround the query in the n-dimensional space; that is, the higher the correlation of the documents with the query the more documents are associated with the query.

3) the set of non-associated documents is partitioned using the standard clustering algorithm, and all loose documents are associated with the nearest partition; this guarantees that every document is included in at least one category; the clusters of documents should be constructed in a similar manner as the cluster of documents used for the two-level search scheme.

In the experimental program, the emphasis has been placed on the various parameters which need to be adjusted since it is necessary in order to validly compare the alternative search procedures either to choose the set of parameters associated with each search scheme so as to maximize effectiveness of the search scheme for the test data base, or to define rules by which it is possible to calculate the value of each parameter for any data base.

D) Test Data Base

The following requirements must be met for the document and query collection to be used to evaluate the effectiveness of the modified versus the normal two-level search:

1) the collection of queries should be real user requests obtained from an actual document retrieval system;

2) the collection of queries should be large enough so that information dense subsets can exist among the queries;

3) relevance judgments should exist for at least a part of the query collection (this provides a control sample of queries which allows the testing of the modified versus normal two-level search scheme for retrieval of relevant documents);

4) the document collection should contain dense areas of information; otherwise, the normal two-level search scheme cannot be efficiently implemented.

5. Actual Comparisons of the Modified versus the Normal Two-Level Searches

In the previous parts, a method of comparing the alternative search procedures was outlined. The method of comparison actually used did not fully follow the suggested method since:

1) a collection of queries created by an actual user population was not available, and further, the available query collection consisted of only 35 queries;

2) the collection of documents available for these queries consisted of only 82 documents from the ADI collection;

3) so many parameters were involved in implementing each search procedure that an adequate appraisal would have required an excessive amount of computer time.

In the framework of this limited data base, the following procedure was actually used:

A) Data Generated for Two-Level Search Algorithm

The standard clustering algorithm was used to partition the collection of 82 documents into 8 clusters, and 10 clusters; each category (cluster) was approximately equal in size. Attempts to divide the document collection into more than 10 clusters were unsuccessful. The number of categories used is not purely arbritrary, since Rocchio [6] proves that if each document has the same probability of being relevant to the query and the categories are approximately equal in size, then the optimum number of categories is equal to $\sqrt{K.82}$, where K is the number of categories which must be searched. If K = 1, then 9 categories should be used so that the sets of 8 and 10 categories are not unreasonable.

B) Data Generated for Modified Two-Level Search Algorithm

Two assumptions are used as the basis for the modified two-level search algorithm:

1) a new query introduced into the system will on a statistical basis be similar to a set of previous queries introduced into the system;

2) a more efficient classification scheme can be constructed if assumption 1) is correct.

It is obvious that 35 queries do not give any indication concerning the truth of the first assumption. In order to carry out the experiment, it was decided to assume the correctness of the first assumption, and to determine instead whether the first assumption implied the second assumption. Two techniques were used to generate a collection of queries which simulated the first assumption, using in each case the 35 queries partitioned into two sets, the first consisting of 25 queries and the second of 10 queries to be used as a control:

1) for each query in the first set, eight random query vectors were generated whose correlations with the initial query were above 0.7. A random query was generated by correlating the initial query with the whole document collection; the vectors representing the two highest correlated documents were summed together with the initial query vector; each concept in this summed vector was then multiplied by a different random number from 0 to 1. This new vector was normalized and correlated with the initial query; if this correlation was greater than 0.7, then this random vector was added to the query collection. This procedure was used until 8 vectors were generated for each query in the first

set; therefore, a collection of 200 queries was constructed to simulate the first assumption. The idea motivating this technique was to produce a query vector which was similar to the initial query vector, but would possibly have different concepts and weights. It was felt that this perturbation of the initial query would simulate a set of different users, phrasing the same type of query.

2) the second collection of queries used to simulate the first assumption consisted of the first set of 25 queries.

The data for the modified two-level search was constructed by considering the two collections of queries described above as collections of previous queries introduced into the system. The following procedures were carried out for both collections of queries:

1) the standard clustering algorithm was used to partition the set of previous queries into sets of 6 and 8 clusters;

2) the subset of associated documents for each query cluster was constructed by associating all those documents which correlated highly with the given query centroid vector; the size of the associated subset of documents depended on the number of queries contained in the given query cluster. Making the size dependent on the magnitude of the document correlations with the centroid vector was also tried, but for the document collection used the associated subsets of documents turned out to have the same size for either procedure. This procedure was then repeated for the 6 and 8 clusters of queries.

3) the non-associated documents resulting from step 2) were clustered into <u>two</u> categories; this was done so that the document collection was partitioned into sets of 8 and 10 clusters. Therefore, the number of categories for the modified and normal two-level search schemes were equal.

By these procedures, 6 different sets of classification vectors and associated documents were constructed:

1) 8 categories, based on clustering documents;
2) 8 categories, based on clustering of 200 random queries;
3) 8 categories, based on clustering of the first 25 queries;
4) 10 categories, based on clustering documents;
5) 10 categories, based on clustering of 200 random queries;
6) 10 categories, based on clustering of the first 25 queries.

C) Experimental Evaluation

The 6 sets of classification vectors together with their associated document subsets are used with the sample collection of 35 queries to compare the search efficiency of the modified two-level search scheme with that of the normal two-level search process.

In order to generate the criteria for search effectiveness, the normal mode of querying is altered; the user's request in this modified querying system consists of a query, a value for the correlation cutoff, and a value for the number of documents to be retrieved. Therefore, the sample collection of queries cannot be used directly as test data (i.e. user requests) in the modified querying system since for each query neither the value for the correlation cut-off, nor the value for the number of documents to be retrieved is specified. In an actual information system, these two parameters vary according to the needs of the particular user (e.g. the set of parameters for high recall differs from those for high precision). Therefore, it is felt that the assignment to these two parameters of

constant values for all queries would bias the conclusions of the experiment. In such a case, the conclusions might only be valid for an information retrieval system where the needs are maximized by the particular set of constants chosen for the experiment. Further, there may be requests consisting of a query together with fixed parameter values which are not satisfied by the test document collection.* In this case, the given query would be useless in the evaluation procedure. Accordingly, it was decided that a systematic variatiation of these two parameters for each query would constitute the best approach, since the effect of varying the two parameters on alternative search schemes could then be observed, and average values could be obtained for the criteria over the entire range of these parameters. Each parameter was in fact varied in the following manner:

1) the value for the cut-off correlation was made to range from 0.2 to 0.6 in increments of 0.1;
2) the value for the number of documents to be retrieved was made to range from 3 to 12 in increments of 3.

In this framework, the data for evaluating the search effectiveness of a given search scheme and a given set of classification vectors with their associated document subsets is generated as follows: for each query ($\overline{q}$) contained in the collection of test queries, a set of 20 search requests is constructed by a systematic variation of the second and third parameters as previously described; the following search requests represented as

---

* A user request is considered as a triplet: ($\overline{q}$, c, n), where $\overline{q}$ is a query index vector, c is the correlation cut-off value, and n is the number of documents to be retrieved; a request is "satisfied" by the given document collection if there exist at least n documents in the document collection whose correlation coefficient with the query $\overline{q}$ is above c.

triplets are then constructed: $(\bar{q}, 0.2, 3)$, ... , $(\bar{q}, 0.2, 12)$, $(\bar{q}, 0.3, 3)$, ..., $(\bar{q}, 0.6, 9)$, $(\bar{q}, 0.6, 12)$. For each search request which is satisfied, the following data are then obtained:

1) M $(\bar{q}, c, n)$ equal to the minimum number of documents which are scanned by the given search scheme in order to "satisfy" the request;

2) P $(\bar{q}, c, n)$ equal to the percentage of documents retrieved by the given search scheme contained in the first n documents retrieved by a full search of the document collection;

3) R $(\bar{q}, c, n)$ equal to the number of relevant documents contained in the set of documents retrieved by the given search scheme over the number of relevant documents contained in the first n documents retrieved by a full search of the document collection.

Let $Q = \{\bar{q}_1, \bar{q}_2, \ldots, \bar{q}_n\}$ be defined as the given collection of n test queries, and let $Q_{co,no} = \{\bar{q}_{i_1}, \bar{q}_{i_2}, \ldots, \bar{q}_{i_{Kco,no}}\}$ be defined as the set of all queries such that $\bar{q}_j \in Q_{co,no} \subset Q$ implies that the request $(\bar{q}_j, co, no)$ is "satisfied"; the data produced from all requests which were satisfied is then condensed in the following manner:

let

$$\bar{M}(co, no) = \frac{\sum_{j=1}^{Kco,no} M(\bar{q}i_j, Co, No)}{Kco,no}$$

$$\bar{P}(co, no) = \frac{\sum_{j=1}^{Kco,no} P(\bar{q}i_j, Co, No)}{Kco,no}$$

$$\bar{R}(co, no) = \frac{\sum_{j=1}^{Kco,no} R(\bar{q}i_j, Co, No)}{Kco,no}$$

$$M_T^{*} = \frac{\sum_{c=0.2 \text{ in steps of } 0.1}^{0.6} \sum_{n=3 \text{ in steps of } 3}^{12} \sum_{j=1}^{Kc,n} M(qi_j,c,n)}{\sum_{c=0.2}^{0.6} \sum_{n=3}^{12} Kc,n}$$

$$P_T = \frac{\sum_{c=0.2}^{0.6} \sum_{n=3}^{12} \sum_{j=1}^{Kc,n} P(qi_j,c,n)}{\sum_{c=0.2}^{0.6} \sum_{n=3}^{12} Kc,n}$$

$$R_T = \frac{\sum_{c=0.2}^{0.6} \sum_{n=3}^{12} \sum_{j=1}^{Kc,n} R(qi_j,c,n)}{\sum_{c=0.2}^{0.6} \sum_{n=3}^{12} Kc,n}$$

The values of M_T, P_T and R_T[**] represent average values for the criteria over the entire range of user needs. This provides a measure of search effectiveness for a given search scheme, and a given set of categories based on a test collection of queries. The values of $\overline{M}$ (co,no), $\overline{P}$ (co,no), $\overline{R}$ (co,no) provide the same type of measure of search efficiency, except that these measures are related to a particular user need (e.g. high recall or high precision, etc.)

---

* M_T could be calculated in the following way:

$$M_T = \frac{\sum_{c=0.2}^{0.6} \sum_{n=3}^{12} \overline{M}(c,n)}{20};$$

However for a limited set of queries this method of calculating is not valid since Kc, n for c, n large will be very small, (i.e. covering not many cases), and therefore the value of $\overline{M}$ (c, n) can fluctuate arbitrarily for such a small sample, so that its value is not a good indicator of the search effectiveness, and thus should not be given an equal weight in the averaging procedure.

** The averaging technique used to calculate these criteria is similar to the procedure used to calculate ranked recall.[6]

This evaluation procedure was programmed in Fortran for the CDC 1604 computer. Two additional evaluation criteria not previously mentioned may also be calculated by this program:

1) $C(\overline{q}, c, n)$ equal to the number of document categories which need to be scanned in order to satisfy the request; the quantities C_T, and $\overline{C}(c,n)$ can then be defined in a manner similar to that used for M_T and $\overline{M}(c,n)$;
2) the average correlation value for the set of test queries with the highest correlating classification vector, the second highest correlating classification vector, etc.

Figures 2 and 3 represent typical output of the evaluation program. The six sets of categories together with their appropriate search schemes are evaluated for search effectiveness using as test collections of queries the first 25 queries, the last 10 queries, and the entire 35 queries. The collection of the first 25 queries is intended to represent a set of queries which is similar to the set of previous queries used to construct the four sets of categories generated by query clustering. The collection of the last 10 queries is intended to represent a collection of new queries to the system which may or may not be similar to the set of previous queries introduced into the system. The entire collection of 35 queries represents a composite collection of queries which provides an overall evaluation of search effectiveness. Table 1 gives the value of the criteria for search effectiveness for each set of categories and test query collections.

QUERY IDENTIFICATION = QA15C

NUMBER OF DOCUMENTS TO BE RETRIEVED = 9
ABOVE CUT-OFF CORRELATION = .300

USING MODIFIED TWO-LEVEL SEARCH BASED ON DISTRIBUTION OF PREVIOUS QUERIES

CORRELATIONS OF QUERY VERSUS CLASSIFICATION VECTORS

CLASSIFICATION VECTOR	QUERY CORRELATIONS
2	.608
7	.418
1	.337
8	.274
6	.186
4	.182
3	.179
5	.149

HAVE SATISFIED SEARCH BY CHECKING 3 CATEGORIES
AND A TOTAL NUMBER OF DOCUMENTS = 29

RETRIEVED DOCUMENTS RANKED BY CORRELATION WITH QUERY

DOCUMENT NUMBER	DOCUMENT IDENT.	CORRELATION
47	1116R	.573
62	809A	.432
11	1117P	.426
37	306PR	.424
1	1104T	.386
51	1115S	.349
43	1206T	.341
27	504IS	.333
12	1113T	.302

Example of Document Retrieval in the Modified Querying System

Figure 2

Modified Two-Level Search Scheme with 8 categories Based on Clustering of 200 Random Queries with First 25 Queries Used as Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$	$\overline{C}(c,n)$
3	.20	25	6.24	.53	.38	1.24
6	.20	23	7.26	.64	.50	1.13
9	.20	19	11.53	.68	.56	1.42
12	.20	18	15.56	.72	.62	1.78
3	.30	23	6.22	.68	.35	1.22
6	.30	15	11.93	.82	.60	1.40
9	.30	9	15.67	.80	.71	1.78
12	.30	7	25.71	.85	.63	2.57
3	.40	10	11.10	.80	.60	1.60
6	.40	5	13.20	.97	.80	1.60
9	.40	1	9.00	.67	.75	1.00
12	.40	1	15.00	.83	.75	1.00
3	.50	2	4.00	.83	1.50	1.00
6	.50	1	8.00	1.00	1.00	1.00
9	.50	0	0	0	0	0
12	.50	0	0	0	0	0
3	.60	1	4.00	.67	2.00	1.00
6	.60	0	0	0	0	0
9	.60	0	0	0	0	0
12	.60	0	0	0	0	0
TOTALS		160	M_T = 10.54	P_T = .70	R_T = .55	C_T = 1.43

Sample Output From Evaluation Program

Figure 3

Table 1: Search Schemes and Categories Evaluated for Search Effectiveness

	SEARCH SCHEME	SET OF CATEGORIES USED	TEST QUERIES	M_R*	P_T	R_T
1.	Normal Two-Level	8 Categories, Based on Clustering Documents	First 25 Queries	20.12	0.65	0.67
2.	Modified Two-Level	8 Categories, Based on Clustering Random Queries	First 25 Queries	16.54	0.70	0.55
3.	Modified Two-Level	8 Categories, Based on Clustering 25 Queries	First 25 Queries	17.25	0.67	0.60
4.	Normal Two-Level	10 Categories, Based on Clustering Documents	First 25 Queries	22.17	0.64	0.66
5.	Modified Two-Level	10 Categories, Based on Clustering Random Queries	First 25 Queries	17.73	0.74	0.63
6.	Modified Two-Level	10 Categories, Based on Clustering 25 Queries	First 25 Queries	18.56	0.72	0.60
7.	Normal Two-Level	8 Categories, Based on Clustering Documents	Last 10 Queries	20.58	0.65	0.37
8.	Modified Two-Level	8 Categories, Based on Clustering Random Queries	Last 10 Queries	23.96	0.66	0.45
9.	Modified Two-Level	8 Categories, Based on Clustering 25 Queries	Last 10 Queries	25.52	0.63	0.57
10.	Normal Two-Level	10 Categories, Based on Clustering Documents	Last 10 Queries	22.38	0.63	0.46
11.	Modified Two-Level	10 Categories, Based on Clustering Random Queries	Last 10 Queries	23.39	0.70	0.52
12.	Modified Two-Level	10 Categories, Based on Clustering 25 Queries	Last 10 Queries	26.70	0.64	0.53
13.	Normal Two-Level	8 Categories, Based on Clustering Documents	Full 35 Queries	20.29	0.67	0.55
14.	Modified Two-Level	8 Categories, Based on Clustering Random Queries	Full 35 Queries	18.75	0.68	0.55
15.	Modified Two-Level	8 Categories, Based on Clustering 25 Queries	Full 35 Queries	19.73	0.65	0.58
16.	Normal Two-Level	10 Categories, Based on Clustering Documents	Full 35 Queries	22.23	0.72	0.57
17.	Modified Two-Level	10 Categories, Based on Clustering Random Queries	Full 35 Queries	19.41	0.70	0.60
18.	Modified Two-Level	10 Categories, Based on Clustering 25 Queries	Full 35 Queries	21.01	0.64	0.60

* $M_R = M_T$ + (the average number of classification vectors which need to be correlated with each query contained in the test collection of queries). The normal two-level search scheme always compares the query against each classification vector so that if 8 categories are used in the normal two-level search, $M_R = M_T + 8$. The calculation of M_R for the modified two-level search is more difficult since it is dependent on the number of categories of associated documents, the number of categories of non-associated documents, and the particular collection of test queries. Consider the example of 8 categories based on the clustering of the first 25 queries, and the first 25 queries as a test collection; the classification vectors for the subsets of associated documents were formed from the first 25 queries so that each of the first 25 queries correlates highly with at least one of these classification vectors. The modified two-level search scheme can then "satisfy" the request by correlating the query with only the set of classification vectors for the subsets of associated documents. Therefore in this example where there were 6 categories of associated documents, $M_R = M_T + 6$. But, if the test collection of queries consisted of the last 10 queries, the set of classification vectors for the non-associated document subsets needs to be correlated with each query in the test collection in order to satisfy the request, and in this case $M_R = M_T + 8$.

Table 2: Average Correlations of First 25 Queries with Classification Vectors

Case	Set of Categories	Average Correlation with Query: Highest Correlating Classification Vector	Average Correlation with Query: Second Highest Correlating Classif. Vector
1.	8 Categories, Based on Clustering Documents	.45	.35
2.	8 Categories, Based on Clustering Random Queries	.71	.36
3.	8 Categories, Based on Clustering 25 Queries	.59	.45
4.	10 Categories, Based on Clustering Documents	.45	.34
5.	10 Categories, Based on Clustering Random Queries	.76	.43
6.	10 Categories, Based on Clustering 25 Queries	.63	.43

Table 3: Revised Search Scheme Evaluation for Test Collection of First 25 Queries Based on Requests which Retrieved at Least 6 Documents

Case	Search Scheme	Set of Categories	M_R	P_T	R_T
1.	Normal Two-Level	8 Categories, Based on Clustering Documents	23.19	.71	.66
2.	Modified Two-Level	8 Categories, Based on Clustering Random Queries	18.77	.74	.60
3.	Modified Two-Level	8 Categories, Based on Clustering 25 Queries	20.02	.73	.61
4.	Normal Two-Level	10 Categories, Based on Clustering Documents	24.68	.68	.63
5.	Modified Two-Level	10 Categories, Based on Clustering Random Queries	20.58	.76	.65
6.	Modified Two-Level	10 Categories, Based on Clustering 25 Queries	21.56	.74	.63

D) Evaluation Results

The improvement in search efficiency by query clustering can be observed in cases 1-6 in Table 1. In all of these cases, the search efficiency as measured by M_R indicates that the modified two-level search based on query clustering is significantly better than the normal two-level search scheme based on document clustering. The reasons for this improvement in search efficiency can be explained by Table 2: the classification vectors of the categories constructed by query clustering are more highly correlated with the test queries; and they more naturally classify the test query to one particular category. This is indicated by the large differences between the first and second highest correlating classification vectors. These results provide an experimental validation of the theoretical advantages of query clustering as illustrated by Figure 1. Unfortunately, the other two criteria P_T, R_T may contradict the general feeling that the higher the query-document correlations (and therefore the larger the value of P_T), the greater the probability of retrieving relevant documents (and therefore the larger the value of R_T). A positive conclusion based on all three criteria for search effectiveness is thus impossible. Still, it is evident that case 5, which is an example of the modified two-level search scheme, is superior to the two examples of the normal two-level search scheme; the values of M_R and P_T for case 5 are much better than for case 1 and case 4, and the differences in the values of R_T for these three cases are small.

The apparent contradiction caused by differences between the values of P_T and R_T can be resolved if the evaluation results are based only on requests which retrieve more than three documents. It appears that for requests which retrieve only 3 documents, high overlap between the first 3

documents retrieved in a full search does not necessarily imply a high recall for the 3 documents as measured by R_T. Table 3 shows the values for M_R, P_T and R_T based on requests which retrieve more than 3 documents. It is felt that this change in the range of the number of documents to be retrieved does not invalidate the conclusions since in an actual information retrieval system, the majority of the user requests retrieve more than 3 documents. The results exhibited in Table 3 clearly indicate that if the collection of new queries introduced into the system is similar to the collection of previous queries introduced into the system, then the modified two-level search scheme is more efficient than the normal two-level search.

The cases 7-12 in Table 1 indicate, as expected, that if a query is not similar to a subset of previous queries, the normal two-level search is more effective than the modified two-level search. It is believed that due to an unanticipated error in the experimental procedures this difference in search effectiveness is unduly increased. The set of non-associated documents was found to be approximately equal in size to the set of associated documents, and was partitioned into only 2 categories. It is felt in retrospect that this was a mistake, and that if the set of non-associated documents had been clustered into 4 categories, the search effectiveness of the modified two-level search scheme for the test collection of the last 10 queries would have been greatly improved. Four categories divides this set of non-associated documents into subsets of documents of the same size as the clusters of associated documents.

The collection of 35 test queries is converted into 231 satisfied requests of which 160 requests were produced from the first 25 queries.

Therefore, approximately 70% of the requests consist of queries which are similar to the set of previous queries introduced into the system. The cases 13-18 indicate that for this type of request distribution the modified two-level search scheme is still more effective than the normal two-level search. In a real user population, a much higher percentage of requests should be similar to the previous requests so that the modified two-level search scheme should be significantly better than the normal two-level search.

6. A New Criterion for Search Effectiveness

It is felt that the use of the three criteria to evaluate search effectiveness is not completely adequate since there is no intuitive procedure to combine the values for each criterion into one composite score which represents overall search effectiveness. This inadequacy in using the three criteria for search effectiveness is demonstrated by the difficulty encountered in evaluating the test results in this experiment. The following is a procedure to calculate the value of a single overall criterion for search effectiveness: for a given test query and for a given search scheme a total ranking is induced on the documents contained in the test document collection; the first document retrieved by the given search scheme for the given test query is ranked number 1, and likewise the n^{th} document retrieved is ranked number n. This procedure is continued until all relevant documents are retrieved. The ranking of the documents relevant to the given test query is determined, and the evaluation measure "ranked recall" [6] is calculated. The average ranked recall for all test queries is then taken as indicative of search

effectiveness.

It is felt that this new criterion based completely on the position of relevant documents is an adequate indicator of relevant search effectiveness of the modified and normal two-level search schemes.

7. Conclusions

The limited test data available, and the restrictions placed on the experiment have not allowed a definitive evaluation of the modified and normal two-level search schemes. It has however been shown that the modified two-level search has potential merit and should be investigated more thoroughly. Hopefully, this report provides an outline for future research into the development of this new search technique.

References

[1] R. E. Bonner, On Some Clustering Techniques, IBM Journal of Research and Development, Vol. 8, No. 1, January 1964.

[2] H. Borko, The Construction of an Empirically Based Mathematically Derived Classification System, SP-585, October 1962.

[3] H. Borko, Research in Computer Based Classification Systems, International Study Conference on Classification Research, 1964.

[4] R. M. Curtice and V. Rosenberg, Optimizing Retrieval Results with Man-Machine Interaction, Center for Information Sciences, Lehigh University.

[5] P. Reisner, Construction of a Growing Thesaurus by Conversational Interaction in a Man-Machine System, in Some Problems in Information Science with Emphasis on Adaptation to Use Through Man-Machine Interaction, editor, M. Kochen, IBM, 1964.

[6] J. J. Rocchio, Document Retrieval System-Optimization and Evaluation, Harvard University Doctoral Thesis, Report ISR-10 to the National Science Foundation, March 1966.

[7] J. J. Rocchio and G. Salton, Information Search Optimization and Interactive Retrieval Techniques, Proceedings of the Fall Joint Computer Conference, Las Vegas, November 1965.

[8] G. Salton, A Document Retrieval System for Man-Machine Interaction, Proceedings of the 19th National Conference, ACM Philadelphia, 1964, pp. L2.3-1 - L2.3-20.

[9] G. Salton, The Evaluation of Automatic Retrieval Procedures - Selected Test Results Using the SMART System, American Documentation, Vol. 16, No. 3, July 1965.

[10] G. Salton, Information Storage and Retrieval, Report ISR-9 to the National Science Foundation.

[11] G.Salton and M. Lesk, The SMART Automatic Document Retrieval System - An Illustration, Communications of the ACM, Vol. 8, No. 6, June 1965.

[12] H. E. Stiles, The Association Factor in Information Retrieval, Journal of the ACM, Vol. 8, No. 2, April 1961.

APPENDIX A

Evaluation Output

Case 1 - The Normal Two-Level Search Scheme with 8 Categories Based on the Clustering of Documents with the First 25 Queries used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$	$\overline{C}(c,n)$
3	.20	25	5.52	.44	.60	1.16
6	.20	23	10.48	.58	.63	1.43
9	.20	19	13.37	.64	.63	1.74
12	.20	18	18.67	.69	.65	2.17
3	.30	23	9.83	.64	.59	1.57
6	.30	15	13.47	.76	.55	1.80
9	.30	9	17.67	.81	.67	2.00
12	.30	7	23.43	.90	.86	2.86
3	.40	10	7.70	.70	.85	1.40
6	.40	5	18.20	.93	.80	2.40
9	.40	1	10.00	.89	1.00	1.00
12	.40	1	19.00	1.00	1.00	2.00
3	.50	2	5.50	.67	1.50	1.00
6	.50	1	8.00	.83	1.33	1.00
9	.50	0	0	0	0	0
12	.50	0	0	0	0	0
3	.60	1	3.00	.33	2.00	1.00
6	.60	0	0	0	0	0
9	.60	0	0	0	0	0
12	.60	0	0	0	0	0
TOTALS		160	M_T = 12.12	P_T = .65	R_T = .67	C_T = 1.67

Case 2 - The Modified Two-Level Search Scheme with 8 Categories Based on the Clustering of 200 Random Queries with the First 25 Queries used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$	$\overline{C}(c,n)$
3	.20	25	6.24	.53	.38	1.24
6	.20	23	7.26	.64	.50	1.13
9	.20	19	11.53	.68	.56	1.42
12	.20	18	15.56	.72	.62	1.78
3	.30	23	6.22	.68	.35	1.22
6	.30	15	11.93	.82	.60	1.40
9	.30	9	15.57	.80	.71	1.78
12	.30	7	25.71	.85	.63	2.57
3	.40	10	11.10	.80	.60	1.60
6	.40	5	13.20	.97	.80	1.60
9	.40	1	9.00	.67	.75	1.00
12	.40	1	15.00	.83	.75	1.00
3	.50	2	4.00	.83	1.50	1.00
6	.50	1	8.00	1.00	1.00	1.00
9	.50	0	0	0	0	0
12	.50	0	0	0	0	0
3	.60	1	4.00	.67	2.00	1.00
6	.60	0	0	0	0	0
9	.60	0	0	0	0	0
12	.60	0	0	0	0	0
TOTALS		160	$M_T = 10.54$	$P_T = .70$	$R_T = .55$	$C_T = 1.43$

Case 3 - The Modified Two-Level Search Scheme with 8 Categories Based on the Clustering of the First 25 Queries with the First 25 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$	$\overline{C}(c,n)$
3	.20	25	4.16	.45	.52	1.04
6	.20	23	8.00	.64	.60	1.89
9	.20	19	13.95	.66	.52	1.42
12	.20	18	16.83	.70	.59	1.44
3	.30	23	6.78	.64	.43	1.13
6	.30	15	16.67	.78	.49	1.73
9	.30	9	17.78	.83	.63	2.00
12	.30	7	18.14	.89	.71	1.86
3	.40	10	14.10	.73	.80	1.80
6	.40	5	14.00	.97	.87	1.60
9	.40	1	10.00	.89	1.00	1.00
12	.40	1	13.00	.92	1.00	1.00
3	.50	2	4.00	.67	1.50	1.00
6	.50	1	6.00	.83	1.33	1.00
9	.50	0	0	0	0	0
12	.50	0	0	0	0	0
3	.60	1	3.00	.33	2.00	1.00
6	.60	0	0	0	0	0
9	.60	0	0	0	0	0
12	.60	0	0	0	0	0
TOTALS		160	M_T = 11.25	P_T = .67	R_T = .60	C_T = 1.37

Case 4 - The Normal Two-Level Search Scheme with 10 Categories Based on the Clustering of Documents with the First 25 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$	$\overline{C}(c,n)$
3	.20	25	7.00	.47	.64	1.40
6	.20	23	9.70	.56	.61	1.48
9	.20	19	12.89	.62	.52	1.58
12	.20	18	16.89	.67	.60	2.28
3	.30	23	10.70	.62	.61	1.65
6	.30	15	14.67	.72	.58	1.87
9	.30	9	19.67	.79	.67	2.56
12	.30	7	23.43	.85	.86	2.71
3	.40	10	7.80	.70	.85	1.30
6	.40	5	13.20	.93	.80	1.80
9	.40	1	10.00	.78	1.00	1.00
12	.40	1	17.00	.83	1.00	2.00
3	.50	2	5.50	.67	1.50	1.00
6	.50	1	8.00	.83	1.33	1.00
9	.50	0	0	0	0	0
12	.50	0	0	0	0	0
3	.60	1	3.00	.33	2.00	1.00
6	.60	0	0	0	0	0
9	.60	0	0	0	0	0
12	.60	0	0	0	0	0
TOTALS		160	M_T = 12.17	P_T = .64	R_T = .66	C_T = 1.73

Case 5 - The Modified Two-Level Search Scheme with 10 Categories Based on the Clustering of 200 Random Queries with the First 25 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$	$\overline{C}(c,n)$
3	.20	25	3.28	.60	.46	1.00
6	.20	23	7.13	.66	.54	1.13
9	.20	19	11.47	.69	.58	1.42
12	.20	18	17.78	.76	.69	1.94
3	.30	23	6.17	.72	.50	1.22
6	.30	15	12.20	.82	.60	1.53
9	.30	9	15.22	.84	.78	2.00
12	.30	7	18.00	.87	.71	2.29
3	.40	10	11.00	.83	.90	1.70
6	.40	5	14.00	.97	.80	2.00
9	.40	1	9.00	.89	1.00	1.00
12	.40	1	12.00	.83	1.00	1.00
3	.50	2	3.00	.83	1.50	1.00
6	.50	1	7.00	.83	1.33	1.00
9	.50	0	0	0	0	0
12	.50	0	0	0	0	0
3	.60	1	3.00	.67	2.00	1.00
6	.60	0	0	0	0	0
9	.60	0	0	0	0	0
12	.60	0	0	0	0	0
TOTALS		160	M_T = 9.93	P_T = .74	R_T = .63	C_T = 1.44

Case 6 - The Modified Two-Level Search Scheme with 10 Categories Based on the Clustering of the First 25 Queries with the First 25 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	M(c,n)	P(c,n)	R(c,n)	C(c,n)
3	.20	25	3.56	.61	.42	1.04
6	.20	23	7.91	.64	.57	1.13
9	.20	19	12.84	.68	.50	1.47
12	.20	18	15.83	.73	.63	1.50
3	.30	23	5.65	.72	.54	1.22
6	.30	15	14.00	.78	.59	1.60
9	.30	9	17.78	.85	.78	2.22
12	.30	7	19.00	.89	.71	2.00
3	.40	10	11.70	.73	.75	1.80
6	.40	5	20.20	.97	.87	2.20
9	.40	1	9.00	1.00	1.00	1.00
12	.40	1	13.00	.92	1.00	1.00
3	.50	2	3.50	.83	1.00	1.00
6	.50	1	6.00	.83	1.33	1.00
9	.50	0	0	0	0	0
12	.50	0	0	0	0	0
3	.60	1	3.00	.67	1.00	1.00
6	.60	0	0	0	0	0
9	.60	0	0	0	0	0
12	.60	0	0	0	0	0
TOTALS		160	M_T = 10.56	P_T = .72	R_T = .60	C_T = 1.42

Case 7 - The Normal Two-Level Search Scheme with 8 Categories Based on the Clustering of Documents with the Last 10 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$
3	.20	10	3.60	.40	.22
6	.20	10	8.00	.53	.31
9	.20	10	13.60	.66	.38
12	.20	10	22.00	.78	.55
3	.30	10	8.10	.60	.27
6	.30	8	16.00	.79	.50
9	.30	3	31.00	.89	.67
12	.30	2	30.50	.96	.50
3	.40	5	4.00	.73	.40
6	.40	2	21.00	.83	0
9	.40	0	0	0	0
12	.40	0	0	0	0
3	.50	1	3.00	.33	0
6	.50	0	0	0	0
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	0	0	0	0
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		71	$M_T = 12.68$	$P_T = .65$	$R_T = .37$

Case 8 - The Modified Two-Level Search Scheme with 8 Categories Based on the Clustering of 200 Random Queries with the Last 10 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$
3	.20	10	3.00	.43	.35
6	.20	10	7.30	.55	.52
9	.20	10	17.70	.60	.42
12	.20	10	27.10	.77	.70
3	.30	10	10.80	.67	.28
6	.30	8	25.88	.77	.41
9	.30	3	26.33	.81	.67
12	.30	2	33.50	.82	1.00
3	.40	5	11.50	.92	.40
6	.40	2	30.00	.80	0
9	.40	0	0	.83	0
12	.40	0	0	0	0
3	.50	1	3.00	0	0
6	.50	0	0	.33	0
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	0	0	0	0
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		71	M_T = 15.96	P_T = .66	R_T = .45

Case 9 - The Modified Two-Level Search Scheme with 8 Categories Based on the Clustering of the First 25 Queries with the Last 10 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$
3	.20	10	3.80	.37	.55
6	.20	10	8.40	.50	.50
9	.20	10	17.50	.58	.52
12	.20	10	31.20	.69	.73
3	.30	10	16.20	.63	.53
6	.30	8	24.50	.79	.51
9	.30	3	30.33	.78	1.00
12	.30	2	46.50	.87	1.00
3	.40	5	8.00	.80	.40
6	.40	2	28.00	.92	.50
9	.40	0	0	0	0
12	.40	0	0	0	0
3	.50	1	4.00	.67	0
6	.50	0	0	0	0
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	0	0	0	0
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		71	M_T = 17.52	P_T = .63	R_T = .57

Case 10 - The Normal Two-Level Search Scheme with 10 Categories Based on the Clustering of Documents with the Last 10 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\bar{M}(c,n)$	$\bar{P}(c,n)$	$\bar{R}(c,n)$
3	.20	10	4.10	.33	.40
6	.20	10	8.30	.55	.34
9	.20	10	13.90	.69	.54
12	.20	10	21.20	.78	.78
3	.30	10	9.10	.50	.48
6	.30	8	11.75	.75	.39
9	.30	3	26.00	.85	.33
12	.30	2	36.50	.92	.50
3	.40	5	8.80	.73	.40
6	.40	2	10.50	.83	0
9	.40	0	0	0	0
12	.40	0	0	0	0
3	.50	1	3.00	.33	0
6	.50	0	0	0	0
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	0	0	0	0
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		71	M_T = 12.38	P_T = .63	R_T = 46

Case 11 - The Modified Two-Level Search Scheme with 10 Categories Based on the Clustering of 200 Random Queries with the Last 10 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$
3	.20	10	3.50	.57	.40
6	.20	10	7.70	.48	.40
9	.20	10	12.20	.64	.54
12	.20	10	23.90	.77	.92
3	.30	10	10.30	.73	.43
6	.30	8	15.63	.81	.47
9	.30	3	33.67	.89	.67
12	.30	2	33.00	.92	1.00
3	.40	5	9.20	.87	.40
6	.40	2	17.00	.92	0
9	.40	0	0	0	0
12	.40	0	0	0	0
3	.50	1	3.00	.67	0
6	.50	0	0	0	0
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	0	0	0	0
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		71	M_T = 13.39	P_T = .70	R_T = .52

Case 12 - The Modified Two-Level Search Scheme with 10 Categories Based on the Clustering of the First 25 Queries with the Last 10 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\bar{M}(c,n)$	$\bar{P}(c,n)$	$\bar{R}(c,n)$
3	.20	10	3.90	.40	.35
6	.20	10	8.10	.50	.52
9	.20	10	17.00	.59	.64
12	.20	10	28.30	.71	.73
3	.30	10	15.50	.63	.43
6	.30	8	24.50	.79	.51
9	.30	3	30.33	.85	1.00
12	.30	2	33.50	.87	1.00
3	.40	5	7.00	.80	.40
6	.40	2	30.50	.83	0
9	.40	0	0	0	0
12	.40	0	0	0	0
3	.50	1	3.00	.67	0
6	.50	0	0	0	0
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	0	0	0	0
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		71	M_T = 16.70	P_T = .64	R_T = .53

Case 13 - The Normal Two-Level Search Scheme with 8 Categories Based on the Clustering of Documents with the Full 35 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	M(c,n)	P(c,n)	R(c,n)
3	.20	35	3.83	.45	.47
6	.20	33	7.79	.61	.58
9	.20	29	15.24	.64	.48
12	.20	28	20.50	.73	.63
3	.30	33	8.00	.65	.39
6	.30	23	19.87	.78	.46
9	.30	12	19.92	.82	.64
12	.30	9	21.56	.90	.78
3	.40	15	13.27	.76	.67
6	.40	7	18.57	.93	.62
9	.40	1	10.00	.89	1.00
12	.40	1	13.00	.92	1.00
3	.50	3	3.57	.56	1.00
6	.50	1	6.00	.83	1.33
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	1	3.00	.33	2.00
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		231	M_T = 12.70	P_T = .67	R_T = .55

Case 14 - The Modified Two-Level Search Scheme with 8 Categories Based on the Clustering of 200 Random Queries with the Full 35 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\bar{M}(c,n)$	$\bar{P}(c,n)$	$\bar{R}(c,n)$
3	.20	35	5.51	.49	.43
6	.20	33	7.61	.60	.50
9	.20	29	13.59	.64	.55
12	.20	28	21.14	.71	.66
3	.30	33	9.24	.67	.40
6	.30	23	16.30	.81	.57
9	.30	12	19.33	.80	.78
12	.30	9	30.22	.85	.71
3	.40	14	5.57	.79	.50
6	.40	7	17.43	.95	.71
9	.40	1	9.00	.67	.75
12	.40	1	15.00	.83	.75
3	.50	3	4.00	.78	1.00
6	.50	1	8.00	1.00	1.00
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	1	4.00	.67	2.00
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		230	M_T = 12.44	P_T = .68	R_T = .55

Case 15 - The Modified Two-Level Search Scheme with 8 Categories Based on the Clustering of the First 25 Queries with the Full 35 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$
3	.20	35	4.97	.43	.49
6	.20	33	9.73	.57	.53
9	.20	29	13.45	.65	.54
12	.20	28	19.86	.72	.62
3	.30	33	9.30	.63	.49
6	.30	23	14.35	.77	.53
9	.30	12	21.00	.83	.67
12	.30	9	25.00	.92	.78
3	.40	15	6.47	.71	.70
6	.40	7	19.00	.90	.57
9	.40	1	10.00	.89	1.00
12	.40	1	19.00	1.00	1.00
3	.50	3	4.67	.56	1.00
6	.50	1	8.00	.83	1.33
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	1	3.00	.33	2.00
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		231	M_T = 12.29	P_T = .65	R_T = .58

Case 16 - The Normal Two-Level Search Scheme with 10 Categories Based on the Clustering of Documents with the Full 35 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$
3	.20	35	3.54	.60	.41
6	.20	33	7.85	.60	.52
9	.20	29	12.62	.67	.51
12	.20	28	18.71	.74	.73
3	.30	33	7.06	.73	.51
6	.30	23	14.57	.79	.55
9	.30	12	21.75	.86	.75
12	.30	9	22.11	.90	.78
3	.40	15	10.87	.78	.63
6	.40	7	19.29	.95	.62
9	.40	1	9.00	1.00	1.00
12	.40	1	13.00	.92	1.00
3	.50	3	3.33	.78	.67
6	.50	1	6.00	.83	1.33
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	1	3.00	.67	1.00
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		231	M_T = 11.43	P_T = .72	R_T = .57

Case 17 - The Modified Two-Level Search Scheme with 10 Categories Based on the Clustering of 200 Random Queries with the Full 35 Queries Used as a Test Collection of Queries

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$
3	.20	35	3.46	.54	.43
6	.20	33	7.42	.61	.54
9	.20	29	13.38	.66	.60
12	.20	28	21.68	.74	.71
3	.30	33	9.00	.70	.48
6	.30	23	16.48	.81	.57
9	.30	12	19.00	.84	.83
12	.30	9	21.44	.87	.78
3	.40	14	5.14	.81	.71
6	.40	7	18.71	.93	.57
9	.40	1	9.00	.89	1.00
12	.40	1	12.00	.83	1.00
3	.50	3	3.00	.78	1.00
6	.50	1	7.00	.83	1.33
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	1	3.00	.67	2.00
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		230	M_T = 11.74	P_T = .70	R_T = .60

Case 18 - The Modified Two-Level Search Scheme with 10 Categories Based on the Clustering of the First 25 Queries with the Full 35 Queries Used as a Test Collection of Queries.

n	c	K(c,n)	$\overline{M}(c,n)$	$\overline{P}(c,n)$	$\overline{R}(c,n)$
3	.20	35	6.17	.43	.57
6	.20	33	9.27	.56	.53
9	.20	29	13.24	.64	.53
12	.20	28	18.43	.71	.66
3	.30	33	10.21	.59	.57
6	.30	23	13.65	.73	.52
9	.30	12	21.25	.81	.58
12	.30	9	26.33	.86	.78
3	.40	15	8.13	.71	.70
6	.40	7	12.43	.90	.57
9	.40	1	10.00	.78	1.00
12	.40	1	17.00	.83	1.00
3	.50	3	4.67	.56	1.00
6	.50	1	8.00	.83	1.33
9	.50	0	0	0	0
12	.50	0	0	0	0
3	.60	1	3.00	.33	2.00
6	.60	0	0	0	0
9	.60	0	0	0	0
12	.60	0	0	0	0
TOTALS		231	M_T = 12.23	P_T = .64	R_T = .60