IX. The Use of Statistical Significance in Relevance Feedback

J. Steven Brown and Paul D. Reilly

Abstract

A new approach to relevance feedback, the statistically significant concept (SSC) approach, is presented; feedback iteration queries are constructed using concepts shown to be statistically significant in differentiating relevant from nonrelevant documents rather than using entire document and query vectors as entities. Three new query types for testing the SSC are presented, and the results of testing these queries are given. The experimental queries are found to be approximately equivalent to Rocchio-type methods in results produced, regardless of the evaluation criterion (including a newly developed frozen exponential ranking factor [FERF]) used, but future study is recommended and courses of investigation are outlined.

1. Introduction

One of the major problems which an information retrieval system must solve is the determination of the correspondence between a given query and the information which the user really wishes to obtain. Often the user supplies a request which is too inaccurate, too brief, or too poorly worded for precise retrieval of the documents relevant to his needs. One method for improving the performance of a document retrieval system is to display items found during a preliminary search of the document files and to ask the user to score these documents as either relevant or non-relevant to his query. The system then generates a new query by combining the information from these judgments and from the known characteristics (words used, ideas expressed, bibliographic entries, etc.) of the documents retrieved. Several algorithms, among them that of J. J. Rocchio (1, 2, 3) and that of R. Crawford and H. Melzer (4), have been developed to address this technique of relevance feedback.

Nearly all of the relevance feedback experimentation to date has utilized the general query update formula cited by Crawford and Melzer (4):

$$(1)\quad Q_{i+1} = \alpha Q_i + \beta Q_0 + \gamma \sum_{i=1}^{N_1} R_i + \delta \sum_{i=1}^{N_2} N_i + \sum_{i=1}^{N_3} w_i \cdot d_i + \sum_{i=1}^{N_4} v_i \cdot c_i \text{ , where } Q_i = \text{query at } i^{th} \text{ ie}$$

where Q_i = query at i^{th} iteration
R_i = relevant documents returned
N_i = nonrelevant documents returned
d_i = vectors of a set of documents considered as "environment"

c_i = vectors of concepts showing imposed relationships

α, β, γ, δ, w_i, v_i = weights

Table 1 details the conditions of some of the experiments. In each approach, a combination of vectors as indivisible entities (that is, the entire vector is used as a term, with no use of only individual parts) is utilized.

The investigation reported in the present paper considers the effect of using statistical tests to select concepts to be manipulated in relevance feedback algorithms. Concepts shown to be significant in differentiating between relevant and nonrelevant documents are used to construct one of three different query forms (Table 2) whose retrieval performance is then tested. The rationale for this statistically significant concept (SSC) approach to relevance feedback is based on the following hypotheses:

(1) The user bases his relevance judgments on only a few of the concepts present in each document;

(2) The small set of concepts which the user employs in his selection more accurately represents the information in which he is interested than does the total set of concepts in the search query;

(3) Those concepts which the user considers important can be determined by a statistical analysis.

	α	β	γ	δ	w_i	v_i
Ide (5)	1	0	1	0	0	0
Riddle, Horwitz, Dietz (6)	0	1	1	0	0	0
Crawford, Melzer (4)	0	0	1	0	0	0
Rocchio (1,2,3)	1	0	$1/N_1$	$-1/N_2$	0	0

Experimental conditions

Table 1

	Positively significant and negatively significant concepts	Nonsignificant concepts
Concept-correlated query	Mean of relevant document concept values	Mean of concept values for all documents
Nonsignificant elements query	0	Remaining elements of original query, if any
Strictly significant query	Mean of relevant document concept values	0

Query Definitions

Table 2

The effort based on these premises is therefore to find a statistical process which will satisfy (3).

The basic SSC approach developed in this study depends on the correlation between the (user-judged) relevance of a document to a given query, and the weights a particular concept has in each retrieved (relevant or nonrelevant) document. One can think of the process as shown in Table 3. As is evident, the document vectors $\underline{d}_i$ are padded with zero weights as necessary so that each vector has the same number of elements; the vectors are then aligned so that the element $d_{ij} = w_{ij}$ represents the weight in document i of concept j (which numbering is determined from an enumeration of all concepts in $C = \{c \mid c \in \underline{d}_i, i = 1, \ldots, n\}$, where C has n elements). The column vector $\underline{S}_i$ then contains as entries w_{ji} the weights of concept c_i in each document retrieved. The relevance vector $\underline{R}$ includes either binary or spectral (graded) relevance judgments for each $\underline{d}_i$, the latter being included to ascertain whether the type of relevance judgment materially affects the correlation.

For each $i = 1, 2, \ldots, m$, $\underline{S}_i$ is correlated against $\underline{R}$ using the Pearson product moment:

Correlation vector $\underline{V}$: | | v_j | |

correlation between $\underline{R}$ and $\underline{S}_j$

$\underline{R}$	$\underline{S}_1$	$\underline{S}_2$	$\underline{S}_3$		$\underline{S}_j$		$\underline{S}_m$	
r_1	w_{11}	w_{12}	w_{13}	...	w_{1j}	...	w_{1m}	$\underline{d}_1$
r_2	w_{21}	w_{22}	w_{23}	...	w_{2j}	...	w_{2m}	$\underline{d}_2$
r_3	w_{31}	w_{32}	w_{33}	...	w_{3j}	...	w_{3m}	$\underline{d}_3$
⋮				⋮				⋮
r_n	w_{n1}	w_{n2}	w_{n3}	...	w_{nj}	...	w_{nm}	$\underline{d}_n$

Column vectors $\underline{S}_i$ = concept vectors

Column vector $\underline{R}$ = relevance vector of user judgments

Vector entry w_{ij} = weight of j^{th} concept in document i

Row vectors $\underline{d}_i$ = document vectors

Correlation between concepts and relevance weights

Table 3

$$(2) \quad P_{xy} = \frac{\sum_{i=1}^{n} (x_i - \mu_1)(y_i - \mu_2)}{\sqrt{\sum_{i=1}^{n} (x_i - \mu_i)^2 \sum_{i=1}^{n} (y_i - \mu_2)^2}} ,$$

$$\text{where } \mu_1 = \frac{1}{n} \sum_{i=1}^{n} x_i , \quad \mu_2 = \frac{1}{n} \sum_{i=1}^{n} y_i$$

A correlation vector $V = (v_1, v_2, \ldots, v_m)$ is then formed using the relation

$$(3) \quad v_i = P_{\underline{R}\,\underline{S}_i} .$$

This coefficient of correlation differs from the cosine value

$$(4) \quad q_{xy} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

in that each occurrence of a vector element x_i or y_i in the cosine coefficient formula is replaced by the term $x_i - \mu_1$ or $y_i - \mu_2$, as appropriate, where μ_i is the mean of the entries of the particular vector. The Person moment thus provides values ranging from -1 to +1 regardless of the signs of the vector elements; the cosine correlation, on the other hand, will be strictly non-negative if all vector entries are non-negative.

As an example of the difference between the two coefficients, one can consider the following two-element vectors:

$$(5) \quad A = (1,\ 10) \qquad B = (1,\ 10) \qquad C = (10,\ 1)$$

The cosine correlation coefficients for these vectors are $q_{AB} = 1.0$ and $q_{AC} = q_{BC} = 0.20$, while the Pearson correlations are $P_{AB} = 1.0$ and $P_{AC} = P_{BC} = -0.98$. The latter value of -0.98 in the Pearson set is indicative of a strong magnitude of association between vectors A and C and vectors B and C; this association is nearly as strong as that between A and B, but the "direction" of association is reversed (that is, a high value of the first component of A implies that the first component of B will also have a high value, but that the first component of C will have a low value, if the associations are assumed to hold among A, B, and C in general). The Pearson moment thus distinguishes the three cases of high negative correlation (in information retrieval, an indication of active user disinterest is a concept), values near zero (a sign of user unconcern regarding a concept), and high positive correlation (an indication of active interest in a concept). These categories correspond to the intuitive ideas of positive and negative relationships as well as provide a basis for a possible extension of the present study to include some variety of negative feedback.

For either correlation method, there exists a test based on Student's t-test (Spiegel [9]) for significance of the difference of the correlation vector component v_i from zero:

$$(6) \quad t = \frac{r_{xy}}{\sqrt{(1 - r_{xy}^2) / (N - 2)}} ,$$

where $N - 2$ represents the number of degrees of freedom of the experiment and r_{xy} represents the correlation.

In the case for which $N = 10$ (10 documents retrieved) one finds that for a one-tailed t-distribution, the following confidence level correlation cutoff values obtain:

Confidence Level	Cutoff x ($\lvert r \rvert > x$)
$p = 0.10$	0.4436
$p = 0.05$	0.5495
$p = 0.01$	0.7159

Correlation Cutoff

Table 4

Cutoff levels of 0.8000, 0.6000, and 0.4000 were chosen for investigation since these values cover the confidence level range fairly well, giving low and high confidence points as well as an average (0.6000) figure.

The above justifications, then, indicate that after the correlations are performed, the cutoff can be used to determine which concepts are significant in the determination of the relevance of the documents retrieved. Those v_i for which $v_i > x$ (x = cutoff value) are termed <u>positively significant concepts</u>, while those for which $v_i < -x$ are termed <u>negatively significant concepts</u>. Other v_i are called <u>non-significant concepts</u>.

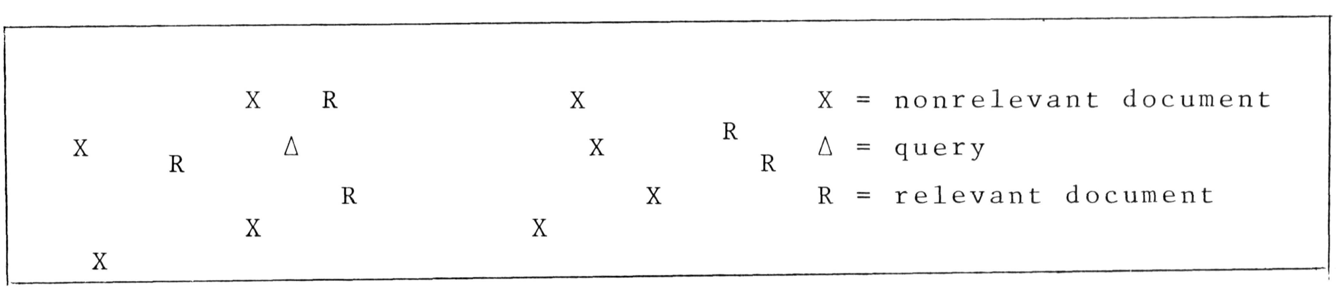
In summary, the key idea underlying the SSC approach to relevance feedback is that document and query vectors are treated as strings of components (the individual concept-weight pairs) rather than as inseparable units. As noted earlier, both the Rocchio and Crawford-Melzer strategies deal with whole vectors, whereas the SSC-oriented methods break the relevance determination into finer levels.

2. Query Construction

The investigation as executed made use of the queries outlined in Table 2, though other combinations of the information obtained by the methods described above are readily apparent. The queries investigated were chosen heuristically as being likely to yield fruitful results.

The first of the three query types, the concept-correlated query, is formed by using the mean of relevant document weights for positively significant concepts and the mean of concept values for all retrieved documents otherwise. This construction is similar to the iteration query proposed by Crawford and Melzer [4] but differs by the component approach described above and by the application of significance tests in determining weights. Clearly, the concept-correlated query may be farther from the original query (in document space) than desirable, and a possible future investigation might examine the attempt to lessen this movement by using the vector composed of the mean of relevant document weights for positively significant concepts and the remaining (i.e., nonsignificant) concepts of the original query.

The nonsignificant elements query is intended as a possible means for approaching the difficulty (pointed out by Ide [7]) of mixed relevant and nonrelevant documents within document space:



Separated Groups of Relevant Items

Figure 1

In this situation, exemplified by the classic sample query requesting information about the aerodynamics of birds, certain concepts of the query (aerodynamics) may greatly overshadow others (birds) in the influencing of the search; the nonsignificant elements query is thus constructed using only those elements of the original query which the significance test has shown to be overshadowed (nonsignificant). It seems feasible that this type of query might be useful in a query-splitting algorithm or in the final stages of an iterative search (in an effort to boost the recall as high as possible).

The third query type, the strictly significant query, is the diametric opposite of the nonsignificant elements query, since the former includes only those concepts of the original query shown to be positively significant (the mean of the concept weight of relevant retrieved documents is used as the entry for each element). The use of this query in iteration will cleary produce a shift in the search toward what the user judged relevant on the previous iteration, and will thus effectively block the retrieval of separated (in document space) new material. As Crawford and Melzer [4] have noted, however, this characteristic has value if the user is highly pleased with the results of the previous retrievals.

The last two methods will in general produce sparse query vectors and hence cannot always be used effectively;

this quality of sparseness (depending on the generality of the concepts and the correlation cutoff level) may cause either the retrieval of a great many documents or the return of a very small number of documents and the disappearance of the vector on successive iterations. A further study, however, may indicate that the set sum (or possibly inter-section in the case of high retrieval from the nonsiginficant elements query) of those documents retrieved under the strictly significant query and under the nonsignificant elements query may produce better results than either method taken alone.

3. Conduct of the Experiment

The study has been carried out using the word form Cranfield 200 Collection [8] and the accompanying 42 queries because this grouping provides a reasonably (and manageably) large number of documents and queries, in view of practical limitations on experimental time and facilities. This collection has the additional advantage of being composed of documents which have been ranked against queries on a five-grade relevance scale; this infor-mation was used rather than the binary judgments in some runs in an effort to test whether the use of finer relevance distinctions would appreciably improve the performance of the queries.

The form of an experimental run, which was conducted within the general context of the SMART information retrieval system, was that of a three-iteration search. The zeroth iteration was a full search, while iterations 1, 2, and 3 were performed using one of the previously described vectors (but the same in all iterations) as the iteration query. Interfacing with the SMART system occurred at four main points: (1) parameter entry, (2) initialization, (3) acquisition and storage of vectors of retrieved documents, and (4) computation of the new query vector. The first three points were covered by trivial mechanical routines, while the fourth point was accomplished by the routine OURCON, which effectively assumed the role of the SMART subprogram MODQUE in creating queries.

In this context, then, several runs of the three query types were made in an effort to determine (1) the effects of the parametric variation of the correlation cutoff level, (2) the effect of the use of spectral relevance judgments as opposed to binary decisions, and (3) the performance of each of the three query types shown in Figure 1.

4. Experimental Results

With regard to the cutoff level, it was observed that for the computing facilities available, the 0.6000 figure (corresponding, as noted in Table 2, to a confidence level of about 0.05) was most suitable regardless of the query type investigated. Experimental runs made using a 0.4000 cutoff in all cases produced queries with several hundred concepts, so that the core storage required for continuation of processing soon exceeded that available. The 0.8000 level, on the other hand, caused the query to shrink noticeably, so that 7 out of the 42 queries vanished entirely during the first iteration (the attempt to construct the first experimental query vector) of the search.

Because of this situation, then, the 0.6000 cutoff was used in all production runs. The query types were checked for performance using both the binary and the spectral relevance schemes (Appendix A contains a summary of the spectral scores), and, as shown in Table 5, no appreciable difference between the types of judgments was detected. Queries 11 and 24 show the most consistent differences in performance, and in each of these cases the binary method provided better final ranking than did the spectral approach. Cases in which the spectral method might be judged superior (Q 13, 16, 28) had differences confined to changes of two or less in the rank of a single document.

Query Type	Number of Differences in Final Ranks of Relevant Documents between Binary and Spectral Schemes	Number of Differences Affecting Order of Presentation of Document (iteration level) to User
Strictly significant	0	0
Concept-correlated	8 (Q 11, 13, 15, 16, 24, 26, 28, 42)	2 (Q 11, 24)
Nonsignificant elements	2 (Q 11, 24)	1 (Q 24)

Performance Difference between
Spectral and Binary Relevance
Table 5

The following graphs of document level recall and precision constitute Figure 2. Figure 2a shows the performance of the Rocchio-type method characterized by equation (7), Figure 2b illustrates the behavior of the strictly significant query, Figure 2c shows the curve for the concept-correlated query, and Figure 2d details the performance of the nonsignificant elements query. All plots are for information obtained from averaged results over 30 queries (excluding queries 6, 9, 12, 14, 21, 23, 25, 29, 32, 33, 35, 36, in which all relevant documents were associated with the Cranfield 200 work form collection.

In the graphs, a solid line denotes performance on the zeroth iteration (full search), a dashed line indicates performance on the first iteration, and a dotted line shows results of the second and third iterations (grouped together). Where two iterations map to the same point, the lower numbered iteration key predominates.

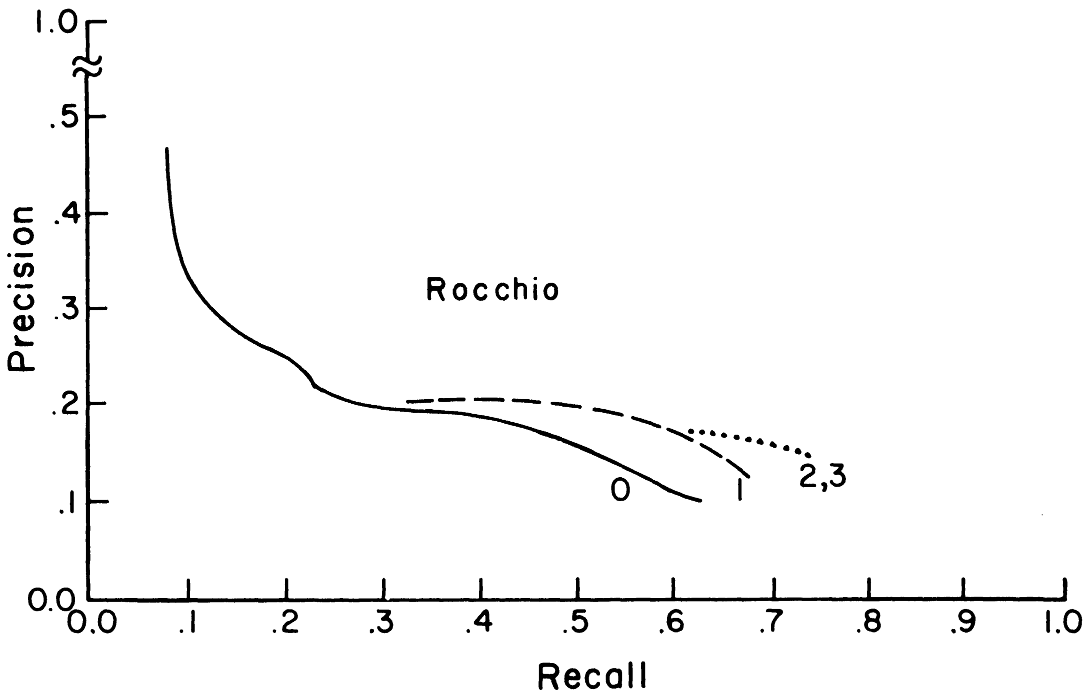


Fig. 2(a)

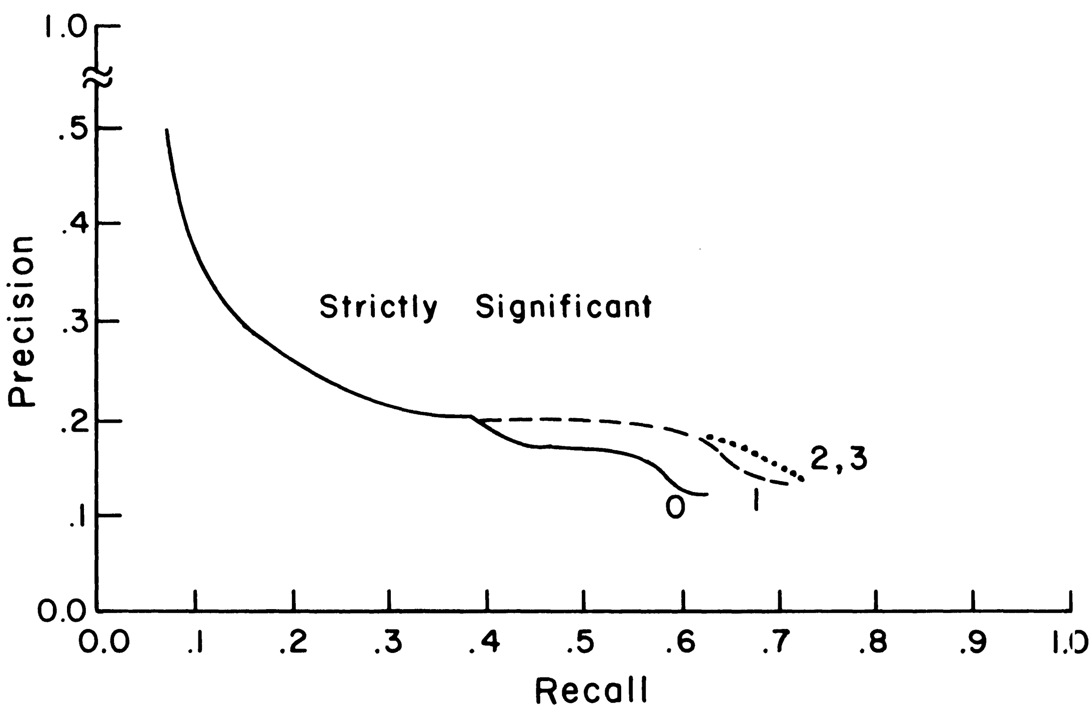


Fig. 2(b)

Recall Precision Graphs

Fig. 2

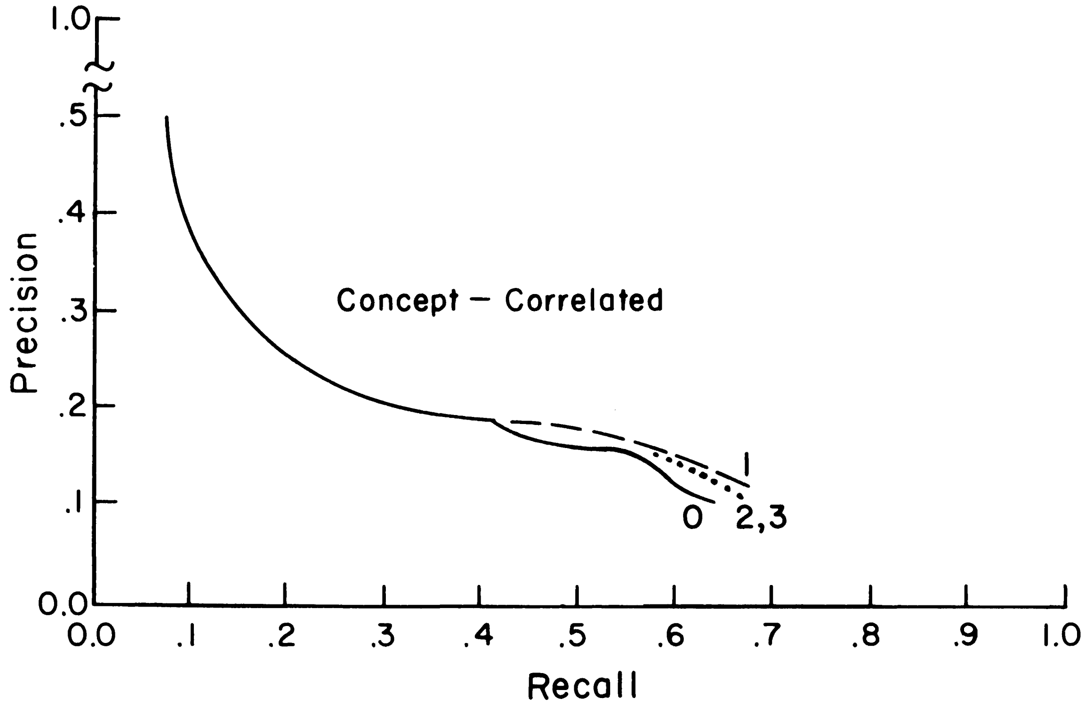


Fig. 2(c)

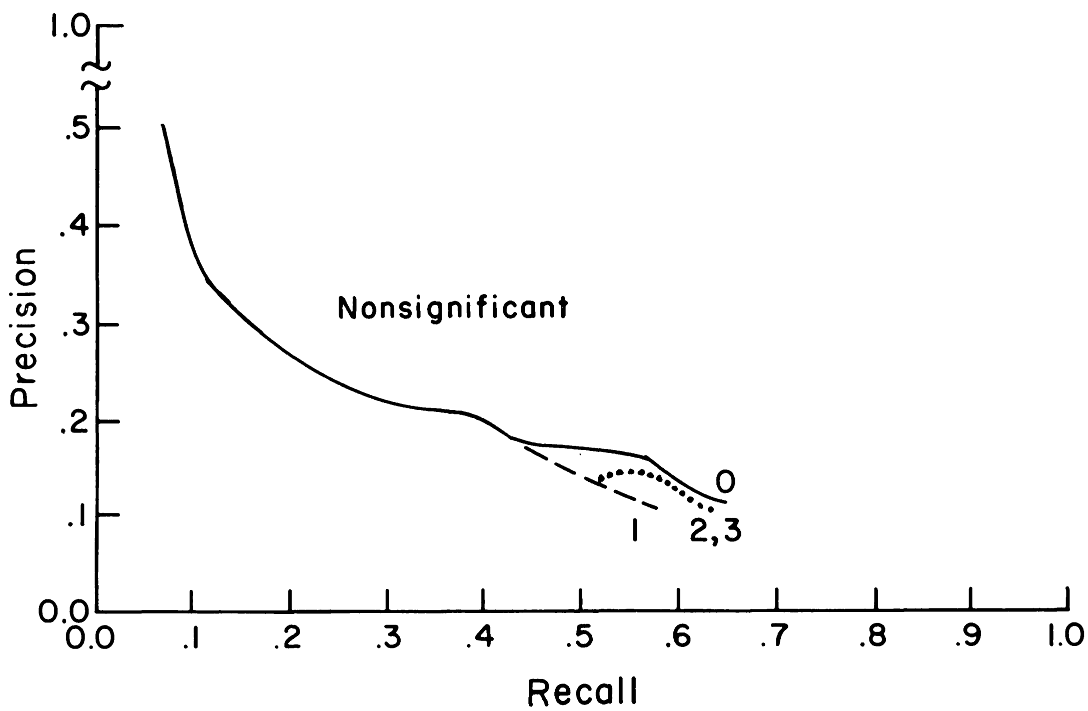


Fig. 2(d)

Recall Precision Graphs

Fig. 2 (contd)

The conclusion that the spectral relevance scheme is not advantageous agrees with what one might expect, since if the matter of binary relevance cannot be consistently judged by different individuals (this fact has been observed by Cleverdon, Mills, and Keen [8]), then the more subjective matter of degrees of relevance would provide a shaky basis indeed for feedback modification.

For the above reasons the general evaluation of the three query types is carried out in the context of a 0.6000 correlation cutoff level and binary relevance grades. In addition, in twelve cases (Q 6,9,12,14,21,23,25,29,32,33, 35,36) all relevant documents were returned to the user on the zeroth iteration (the full search), and, in accordance with methods used by other investigators (e.g., Ide [7]), these queries are excluded from the evaluation.

Plots of the document level recall-precision curves are shown in Figures 2b,2c, and 2d for the strictly significant query, the concept-correlated query, and the nonsignificant elements query, respectively. The curves are generally lower than the curve for a Rocchio-type strategy with similar feedback (Figure 2a):

$$(7) \quad Q_{i+1} = Q_i + \sum_{i=1}^{N_1} R_i \qquad (\alpha=1,\gamma=1,\ \text{all other coefficients}=0\ \text{in equation}(1))$$

Three points should, however, be noted:

(1) The experiment was structured so that if at any time an iteration query failed to return any new relevant documents in the group of ten documents shown to the user, that query was discarded and the original query retrieved and used in its place until either the iteration count was satisfied or another relevant document was retrieved.

(2) If the original full search retrieved no relevant documents before position j ($j > 10$), the original query continued to be used until position j was reached in the return to the user; thus the number of iterations in which the experimental query was used was reduced.

(3) The search comparisons were accomplished using a cosine correlation between query and document vectors taken as entities. A procedure more parallel to the SSC approach might have used a concept-wise correlation dealing only with those concepts appearing in the query.

The first characteristic leads to what might be called the "roller coaster" effect, in which post-search analysis shows that previously unfetched relevant documents moved up sharply during an iteration of the experimental query, but were lost when a return was made to the original query. Examples of this effect are shown in Table 6.

Rank	Query 18 Iteration 0*	1	2*	3*
1	96R	96R	96R	96R
2	140	140	140	140
3	97R	97R	97R	97R
4	199R	199R	199R	199R
5	46	46	46	46
6	64	64	64	64
7	68	68	68	68
8	52	52	52	52
9	47	47	47	47
10	42	42	42	42
11	141	90	90	90
12	108	119	119	119
13	49	120	120	120
14	178	117	117	117
15	161	18	18	18
16	21	27	27	27
17	85	95	95	95
18	171	99	99	99
19	40	194	194	194
20	100	193	193	193
21	10	153	141	141
22	44	89	108	108
23	61	104	49	49
24	54	72	178	178
25	112	155R	161	161
26	187	141	21	21
27	55	121	85	85
28	181	173	171	171
29	157	134	40	40
30	143	93	100	100

* - Original query used

Rank	Query 22 Iteration 0*	1	2*	3*
1	163	163	163	163
2	179	179	179	179
3	167	167	167	167
4	200	200	200	200
5	112	112	112	112
6	130R	130R	130R	130R
7	150	150	150	150
8	125	125	125	125
9	166	166	166	166
10	164	164	164	164
11	14	10	10	10
12	57	137	137	137
13	184	129	129	129
14	58	14	14	14
15	31	142	142	142
16	42	160	160	160
17	102	106	106	106
18	30	136	136	136
19	198	143	143	143
20	189	135	135	135
21	109	140	57	57
22	147	108	184	184
23	39	107	58	58
24	142	139	31	31
25	32	11	42	42
26	60	153	102	102
27	178	24	30	30
28	145	83	198	198
29	103	184	189	189
30	129	149	109	109
55		128R		
69	128R			
74			128R	
75				128R
110		131R		
128	127R			
132		127R	127R	
134				127R
187	131R			
188			131R	
189				131R

"Roller Coaster" Effect from Discarding Experimental Iteration Query
(Strictly significant query, binary judgments, 0.6000 cutoff)

Table 6

It is not clear, however, that the best approach is to continue using the experimental iteration query when it fails to draw any new relevant documents into the group seen by the user. Table 7 shows that the same experimental query may perform well in one case but may actually worsen the results that the original query alone would have obtained in another case. Undoubtedly, part of this behavior can be explained by the fact that the query types tested in this study (particularly the strictly significant query) move the query vector decidedly toward previously retrieved documents, but more investigation into query characteristics and into the reasons why there exist groups of requests which do well under one iterative scheme and not under another is necessary before any comprehensive conclusion can be drawn as to the best path to take in resolving the problem expressed in (1) above.

The second point shows that the results are biased somewhat by the fact that in several cases (for example, Q 10, 11, 20, 26) the experimental query was first used on iteration 2. For query 1 the results were outstandingly bad in all efforts -- the original query had a zero correlation with all relevant documents, and no experimental method was ever applied at all. For future studies, a wiser action in the situation in which no relevant documents

Rank	Query 5 Iteration 0*	1	2	3*
1	59R	59R	59R	59R
2	162	162	162	162
3	197	197	197	197
4	58R	58R	58R	58R
5	160	160	160	160
6	29	29	29	29
7	182	182	182	182
8	189	189	189	189
9	185	185	185	185
10	184	184	184	184
11	60R	13R	13R	13R
12	165	89	89	89
13	115	141	141	141
14	150	60R	60R	60R
15	9	200R	200R	200R
16	191	21	21	21
17	10	15	15	15
18	198	27	27	27
19	164	56	56	56
20	11	126	126	126
21	156	172	123	123
22	139	176	167	167
23	168	171	122	122
24	92	139	109	109
25	28	44	94	94
26	16	30	121	121
27	180	123	18	18
28	57	4	163	163
29	200R	140	165	165
30	90	147	35	35
112		8R		
133	13R			
140	8R			
144				8R

* - Original query used

Rank	Query 7 Iteration 0*	1	2	3*
1	41R	41R	41R	41R
2	90R	90R	90R	90R
3	11	11	11	11
4	60	60	60	60
5	45	45	45	45
6	76	76	76	76
7	160	160	160	160
8	111	111	111	111
9	100	100	100	100
10	176	176	176	176
11	185	95R	95R	95R
12	133	91	91	91
13	192	93	93	93
14	159	36	36	36
15	117	199	199	199
16	110	64	64	64
17	156	173	173	173
18	71	32	32	32
19	83	155	155	155
20	29	96	96	96
21	132	94	117	117
22	158	193	94	94
23	195	43	119	119
24	198	33	103	103
25	150	140	192	192
26	154	116	195	195
27	114	99	92	92
28	199	50	153	153
29	23	23	121	121
30	184	103	120	120
34		42R		
35	95R			
38	72R			
51				72R
65		72R		
69	42R			
79				42R
80			72R	
102			42R	

Advantageous Effects from Discarding Experimental Iteration Query
(Strictly significant query, binary judgments, 0.6000 cutoff)

Table 7

are returned would probably be to utilize a variant of equation (1), perhaps with $\alpha = 1$, $\gamma = 1$, $\delta = 1$, and all other coefficients = 0, to move the query in document space. Alternatively, the inclusion of a negative feedback strategy into the experimental queries, through some equation such as

(8) $Q_{i+1} = Q_e + c\,N$, where N is a vector (possibly with weighted elements) of negatively significant concepts, and where Q_e represents the experimental query of type e,

might be effective, where Q_e could be replaced by Q_i or Q_0 if no relevant documents were returned.

With regard to the performance of specific query types, it was discovered that in all of the 30 user queries in the analysis, the nonsignificant elements query produced final rankings lower than either the strictly significant, the concept-correlated, or the Rocchio-type (formulated as in equation (7)) queries, thus indicating either that the type query for which it is effective is not present in this collection or that the method is generally inapplicable. It is, of course, impossible to draw a general conclusion answering this question from the small request sample involved in the present experiment, but the implication that this query may not be particularly useful (at least when used alone) could follow from either case.

One should note, however, that the experimental non-significant elements queries (first iteration) contained an average of 8.0 concepts, as compared to an average of 9.1 concepts for the original queries. This finding corroborates hypothesis (2) of section 1, which states that only a very few ideas of the original request are really important in determining the user's needs. This last conclusion is noteworthy for its possible application to the interpretation of an original natural language request, since it may imply that a detailed analysis of the query is not necessary because some quick method might be developed to abstract the discriminatory ideas.

The strictly significant query performed in general very similarly to the Rocchio-type query of equation (7). Table 8 details examples in which either method surpassed the other, thus lending support to the feeling that perhaps a real key to more successful retrieval is the development of a strategy by which queries can be classified into groups for which a particular method is appropriate. One should note that experimental results confirm pre-test projections in that the strictly significant query moved decidedly closer to the previously retrieved documents, so that in some cases relevant documents were actually pushed away from retrieval (Table 7).

	Strictly Significant Query		Concept-Correlated Query		Rocchio-Type Method	
Query 3	10	32R	10	32R	10	32R
(Strictly	14	33R	16	30R	18	30R
Significant)*	15	30R	18	33R	21	4R
	20	4R	26	4R	22	57R
	21	57R	27	31R	24	31R
	22	31R	32	57R	124	33R
Query 5	1	59R	1	59R	1	59R
(Strictly	4	58R	4	58R	4	58R
significant)	11	13R	15	200R	12	200R
	14	60R	32	60R	17	60R
	15	200R	134	13R	56	13R
	144	8R	141	8R	82	8R
Query 11	12	92R	12	92R	12	92R
(Concept	14	45R	14	45R	14	45R
correlated)	40	16R	40	16R	61	44R
	119	44R	115	44R	72	16R
Query 17	5	94R	5	94R	5	94R
(Rocchio)	22	90R	22	90R	20	95R
	24	93R	24	93R	21	91R
	32	91R	32	91R	22	90R
	33	95R	38	95R	25	93R
Query 18	1	96R	1	96R	1	96R
(Rocchio)	3	97R	3	97R	3	97R
	4	199R	4	199R	4	199R
					18	155R
Query 39	1	154R	1	154R	1	154R
(Strictly	3	17R	3	17R	3	17R
significant)	14	136R	15	135R	12	135R
	16	135R	51	157R	20	157R
	20	157R	140	136R	39	136R

* - method judged best by FERF criterion over three iterations (Section 3)

All ranks less than 10 were set by the initial full search.

Final Ranks of Relevant Documents for Selected Queries

Table 8

The concept-correlated query followed the same general pattern as the Rocchio-type method, but generally performed noticeably worse than either the Rocchio query or the strictly significant query. If rankings are made using the FERF coefficient (defined below), the concept-correlated query never surpasses the Rocchio method (the two are tied on queries 16,27,28 and 31) and surpasses the strictly significant query only on queries 11, 16,26,27,28, and 31 . In addition, for some queries, such as queries 5 and 39 (Table 8), the concept-correlated method performs considerably worse than does either the Rocchio or the strictly significant type. The most probable explanation for this behavior is that the noise introduced by the entries for nonsignificant concepts (Table 2) is adversely affecting the discrimination of the search.

Document level recall-precision graphs (Figure 2) show that for the 30 queries analyzed, the Rocchio-type method provides a curve which is slighly higher than that of the nonsignificant elements query and the concept-correlated query, and about the same pattern as that of the strictly significant query. It is difficult to explain why the three experimental queries lie so close together in performance unless one concludes that the SSC approach as used in this investigation (though perhaps different results would be found if the three points raised above were resolved) contributes no information to the search that a simple Rocchio method does not also produce.

In an effort to gain a more solid quantitative measure of the performance of an iteration method in a frozen feedback situation (in which documents retrieved have their ranks locked, so that the highest ranking a document can receive on iteration i is $i * N$, where N documents are returned to the user on each iteration), the <u>frozen exponential ranking factor</u> (FERF) has been developed:

$$(9) \quad g_r = T - \sum_{j=0}^{r-1} n_j \quad , \quad r = 1,2,\ldots,i$$

$$(10) \quad f_r = \begin{cases} 0 & \text{if } gr=0 \\ n_r/g_r & \text{otherwise} \end{cases}$$

$$(11) \quad P = \text{FERF} = \sum_{j=0}^{i-1} (10 ** (i-j)) * f_{j+1} \ ,$$

where T = total number of documents relevant to a query

n_k = number of relevant documents retrieved on the k^{th} iteration

i = number of iterations (not counting initial full search) performed

The quantity p, which will always lie in the range [0, 10 ** i], has been introduced as a possible answer to the evaluation problem pointed out by Hall and Weiderman [10]. The FERF is not affected by the number of relevant documents retrieved on the full search (provided all are not found, in which case the evaluation breaks down), and it does assign a higher coefficient to a method which promptly retrieves new material than to a method which retrieves the same material on a later iteration. Furthermore, the FERF is in some sense "normalized" since it is independent of the number of documents relevant to a query.

One should note that the FERF is a rather gross measure of desirability in that it makes no evaluation of rank within an iteration group shown to the user. This limitation, however, is not one of major consequence since the user will presumably examine the entire group returned in any case.

Goodness of result has for these reasons been associated directly with the magnitude of the FERF (an assignment which is intuitively pleasing, as seen in the example below). An illustration of the FERF is given in Table 9.

Following these ideas, the following rankings for the various types of experimental queries in this study obtain

Global conditions: Query Q (notation identical to that in body of the paper)
$i = 2$
$N = 5$
$T = 7$

	Method A			Method B			Method C		
	1			1			1		
	2	R*		2	R		2	R	
Full Search	3			3			3		
	4	R		4	R		4	R	
	5			5			5		
	6	R		6	R		6		
	7		$g_1 = 5$	7		$g_1 = 5$	7		$g_1 = 5$
Iteration 1	8	R		8			8	R	
	9	R	$f_1 = 0.6$	9		$f_1 = 0.4$	9	R	$f_1 = 0.6$
	10			10	R		10	R	
	11			11	R		11		
	12		$g_2 = 2$	12	R	$g_2 = 3$	12		$g_2 = 2$
Iteration 2	13			13	R		13		
	14		$f_2 = 0.0$	14		$f_2 = 1.0$	14		$f_2 = 0.5$
	15			15			15	R	
			FERF = 60			FERF = 50			FERF = 65

* - indicates position of a relevant document

An Illustration of the FERF Approach

Table 9

when averages over the 30 user queries are taken:

Query Type	FERF
Rocchio (equation (7))	547
Strictly significant	475
Concept-correlated	398
Nonsignificant elements	291

Overall Significance Evaluation

Table 10

This measure also shows that the experimental queries do not surpass the simple Rocchio-type query performance.

The SSC method, as tested in this investigation, does require a sizeable amount of time beyond that necessary for an ordinary Rocchio-type relevance feedback search. The information fetches, significance calculations, and query construction increase the machine time of a search by approximately 15% and the time (and effort) required from the user to judge the relevance of ten documents (a quantity which is probably necessary to provide a defensible base for the statistical calculations) is markedly greater than in unembellished methods. Consequently, the SSC approach must be proved capable of producing substantially better results than do existing strategies before the increased resource expenditure necessary to utilize the SSC ideas can be justified.

5. Conclusions and Recommendations

Although the investigations conducted with the queries of Table 2 have not been outstandingly successful in obtaining better methods of relevance feedback, the authors feel that because of the impossibility of testing all aspects of as broad a concept as the SSC feedback approach in a single rather limited experiment, some of the ideas on which the present study is based should be further investigated before they are dropped from consideration. In particular, the approach of treating documents and queries as strings of concept beads which can be broken apart, rather than as indivisible bars which must be added, subtracted, and weighted as entities seems to have value because it allows the investigator to be more selective in filtering out the noise introduced by irrelevant information contained in parts of a document or query vector. The use of statistical tests to ascertain those concepts important in distinguishing relevant from nonrelevant documents should be investigated further, and additional query types should be developed, perhaps along the lines suggested previously, in which the characteristics of the SSC approach can be fully exploited. For example, one could investigate a modified concept-correlated query in which positively significant concepts are entered with the mean of the weights in relevant documents and each remaining (unused) concept of the original (or i^{th}) query is entered with its weight unchanged. Similarly, means of resolution of the situation in which the iteration query retrieves no further relevant documents for a particular iteration and means for introducing negative feedback into the SSC approach should be considered.

Another possible area of future research is the extension of the concept-wise procedure to the actual retrieval of documents, as outlined in the third evaluative point mentioned in Section 4. Further work could also be done in the area of checking correlation cutoff levels; it is now known, for instance, that for the type of feedback reported here that 0.4000 and 0.8000 are beyond the range of workable levels, but the effect of varying the cutoff from 0.6000 to a lesser degree has not been studied.

REFERENCES

[1] Rocchio, J.J. "Relevance Feedback in Information Retrieval", Scientific Report No. ISR-9 to the National Science Foundation, Section III, Harvard Computation Laboratory, August 1965.

[2] Rocchio, J.J. "Document Retrieval System Optimization and Evaluation", Harvard University Doctoral Thesis, Scientific Report No. ISR-10 to the National Science Foundation, Harvard Computation Laboratory, March 1966.

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

[4] Crawford, R.G., Melzer, H.Z. "The Use of Relevant Documents Instead of Queries in Relevance Feedback", Scientific Report No. ISR-14 to the National Science Foundation, Section XIII, Department of Computer Science, Cornell University, October 1968.

[5] Riddle, W., Horwitz, T., Dietz, R. "Relevance Feedback in an Information Retrieval System", Scientific Report No. ISR-11 to the National Science Foundation, Section VI, Department of Computer Science, Cornell University, June 1966.

[6] Ide, E. "User Interaction with an Automated Retrieval System", Scientific Report No. ISR-12 to the National Science Foundation, Section XII, Department of Computer Science, Cornell University, June 1967.

[7] Ide, E. "Relevance Feedback in an Automatic Document Retrieval System", Cornell University Master of Science Thesis, Scientific Report No. ISR-15 to the National Science Foundation, January 1969.

[8] Cleverdon, C., Mills, J., Keen, M. ASLIB Cranfield Research Project - Factors Determining the Performance of Indexing Systems, Cranfield, 1966.

REFERENCES (Continued)

[9] Spiegel, M.R. Statistics, New York, Schaum Publishing Co., 1961 (p. 247).

[10] Hall, H., Weiderman, N. "The Evaluation Problem in Relevance Feedback Systems", Scientific Report No. ISR-12 to the National Science Foundation, Section XII, Department of Computer Science, Cornell University, June 1967.

APPENDIX A

Origins of the Spectral Relevance Judgments for the Cranfield 200 Document Collection

Method of Retrieving Documents

(Abstracted from Cleverdon, Mills, and Keen [8], Vol. 1, p.79)

Stage 1: Authors of documents in the collection construct search questions (queries) and make a relevance assessment of items listed in the bibliographic citations of their own documents which are included in the collection.

Stage 2: Using the document collection and questions from Stage 1, technically competent people examine every document in relation to every question to find any additional (to the bibliographic citations noted in Stage 1) relevant documents.

Stage 3: The document authors receive the additional documents produced by Stage 2 and make a final assessment of relevance.

APPENDIX A (Continued)

Method of Marking Relevance

(Abstracted from Cleverdon, Mills, and Keen [8], Vol. 2, p.123; codes changed to conform to the usage in the present experiment)

Grade 4: References which are a complete answer to the question.

Grade 3: References which are of a high degree of relevance, the lack of which would have made the research (to answer the query) impracticable or would have resulted in a considerable amount of extra work.

Grade 2: References which are useful, either as general background to the work or as suggesting methods of tackling certain aspects of the work.

Grade 1: References which are of minimum interest (for example, those that have been included from a historical viewpoint).

Grade 0: References which are of no interest.