IX. Word-Word Associations in Document Retrieval Systems

M. E. Lesk

1. Introduction

Word normalization procedures in document retrieval systems are traditionally based on manually constructed thesauruses and term lists. Recently, automatic methods dependent on statistical co-occurrence of words have been proposed for the determination of word meanings and the selection of synonymous words, and it has been asserted that the use of such word-occurrence statistics can substitute for thesauruses in retrieval systems. [1,2]

Word-association procedures can be investigated through the SMART automatic document retrieval system, which is capable of simulating a wide variety of proposed computerized text analysis systems in an experimental retrieval environment. [3,4] The SMART system includes methods for automatic processing of text and questions, and for the evaluation of the test results using a variety of performance measures. Existing test collections and dictionaries are used to analyze and evaluate the performance of association procedures for document retrieval.

2. Method

In the SMART retrieval programs, documents are translated into "concept vectors", consisting of a list of concepts with attached weights. Each concept represents a piece of information found in the text by the analysis routines, and the weight reflects the number of times the concept was found and the importance attached to it. The concepts may represent words, groups of synonymous words, phrases, or any other indications reflecting the content of documents. In a word matching system, for example, each English stem is a concept, and the number of occurrences of a stem is its weight. The concept vector then represents a frequency list of the words in the text or query (with suffixes removed). Retrieval tests are performed by matching queries against documents to find the documents with the most similar concept vectors.

To simulate a word-word association process, the concept vectors associated with the requests and documents are augmented by concepts found to be related to the original concepts. The association procedures construct a list of word pairs which are strongly associated, and for each word in the concept vector of a document, all words paired with it are added to the concept vector of the document. The expanded concept vector is used for retrieval in exactly the same fashion as the original concept vector, and the results are compared.

Related word pairs are determined by the following algorithm. For each word in the document collection, a list of the documents in which the word has occurred is compiled and the frequency of the word is noted. For each pair of words, these lists are compared and a measure of similarity between the two concepts is then evaluated. The normal measure of similarity is the "cosine" correlation, defined by

$$r_{ij} = \sum_k w_{ik} w_{jk} \Big/ \sqrt{\sum_k w_{ik}^2 \cdot \sum_k w_{jk}^2} \quad ,$$

where w_{ik} is the weight of word i in document k, and r_{ij} is the correlation between concept i and concept j. Alternatively, the "overlap correlation" may be used; it is defined as

$$r_{ij} = \sum_k \min(w_{ik}, w_{jk}) \Big/ \min(\sum_k w_{ik}, \sum_k w_{jk}) \quad .$$

An example of these correlation procedures is given in Fig. 1. All pairs of words, in which the correlation exceeds a previously set cutoff, are used as associated pairs. These pairs are then employed in the document vector expansion procedure.

Many options are available in this procedure. Either correlation method may be used; the cutoff may be adjusted arbitrarily; the procedure may be iterated with the word similarities measured by the correlation of their lists of related words as determined by the previous iteration; the weights with which the new words are added to the concept vector may be changed; and words occurring outside specified frequency ranges may be omitted from the procedure.

Experiments were performed on three document collections, all used for many SMART experiments. The Cranfield collection, consisting of 200 abstracts in aeronautics collected by the Aslib-Cranfield project in England, is used for most of the investigation. Evaluation is based on a set of 42 actual research questions, with relevance judgments obtained from the researcher himself. The other collections used are the IRE collection, consisting of about 780 abstracts in computer science, and the ADI collection of 82 short papers in documentation. Prepared questions are available for these collections with relevance judgments made by the authors.

Document	1	2	3	4	5
Contains	cat dog fish mouse	cat lion mouse bird	cat bear tiger	dog lion wolf	dog bear mole

a) Document - Term Assignment

Term	cat	dog	fish	mouse	lion	bird	bear	tiger	wolf	mole
Occurs in	1,2,3	1,4,5	1	1,2	2,4	2	3,5	3	4	5

b) Term - Document Assignment

	Cosine	Overlap
$r_{cat, dog}$	$(1+0+0)/\sqrt{(1+1+1)(1+1+1)}$ $= 1/3 = 0.33$	$1/3 = 0.33$
$r_{cat, mouse}$	$(1+1)/3.2$ $= 2/\sqrt{6} = 0.82$	$1/2 = 0.5$

c) Computations of Association

Example of Concept-Concept Association Procedure

Fig. 1

For a cutoff of 0.45, "cat" would be found related to "mouse" but not to "dog". The vector of document 3, after expansion, would then include "cat, bear, tiger, mouse". If the cutoff were 0.6, and the correlation mode were overlap, "cat" and "mouse" would also be found to be unrelated.

The major procedures used for evaluation in the SMART system are described elsewhere. [3,4] They are the recall-precision curve, and four global measures: rank recall, log precision, normalized recall, and normalized precision. The measures vary from 0 to 1, with 0 representing the worst possible performance and 1 representing perfect performance. These measures all reflect both recall and precision, requiring both perfect recall and perfect precision to produce a measure of 1, but the rank recall and normalized recall measures both reflect recall more than precision, while the log and normalized precision reflect precision more strongly than recall. The "quasi-Cleverdon" recall-precision curves shown here are averaged recall-precision curves over the set of 42 requests.

3. Results

Table 1 shows the distribution of association pairs as a function of word frequency, with a cosine correlation at a cutoff of .6. It is seen that the largest number of correlations occur for words of very low frequency, frequencies 1 and 2. With the correlation measure used, it is very easy for low frequency words to co-occur significantly, since, if two words of frequency 1 occur in the same document they will always have a correlation of 1.0. With a collection size of 200 documents, in which 1179 words occur only once, one may expect over 7000 correlations above cutoff of words of frequency 1 with other words of frequency 1 purely on a random basis. If the words of frequency 2 are also considered, the total number of random correlations above .6 would be expected to be about 12000. It is clear therefore that the 18000 correlations observed do not actually

Frequency of One Word of the Pair	Number of Words of That Frequency	Number of Pairs	Average Number of Pairs Per Word	Approx. Weight in Text Expansion	Percent Weight in Expansion
1	1179	18894	16.0	18900	17
2	382	6924	18.1	13800	12
3	199	1855	9.3	5600	5
4	126	1003	8.0	4000	4
5	103	957	9.3	4800	4
6	83	588	7.1	3500	3
7	61	554	9.1	3800	3
8	55	416	7.6	3300	3
9	41	254	6.2	2300	2
10	34	229	6.7	2300	2
11-14	87	819	9.4	9800	9
15-19	54	293	5.4	5300	5
20-29	86	481	5.6	12000	11
30-39	43	87	2.0	3000	3
40-49	25	101	4.0	4500	4
50-59	18	105	5.8	5800	5
60-69	13	32	2.5	2100	2
70-79	8	3	0.4	200	0
80-89	7	34	4.9	2900	3
90-99	6	2	0.3	200	0
100-124	6	5	0.8	600	1
125-149	5	39	7.8	5400	5
150-174	1	0	0.0	0	0
175-199	2	1	0.5	200	0
200+	4	4	1.0	1000	1
all	2628	28680	10.91	111500	100

Word-Word Associations Tabulated by Word Frequency

Table 1

represent significant data but are largely chance occurrences. Since these correlations represent the major part of expanded document vectors, they will perturb the run. We have, therefore, adopted the expedient of removing all correlations involving a word occurring fewer than three times from the runs, thus eliminating most of the chance correlations. The expected number of chance correlations between words occurring three times is only one or two.

Complete elimination of chance correlations requires the removal, not only of the words occurring very few times, but also the words occurring many times. Any two words which occur in more than half the documents in the collection, for example, are clearly likely to have a high correlation; the expected chance correlation between two words, each occurring in 100 documents, is about .5, which is quite close to the cutoff. The expected correlation between two words occurring in every document is almost certain to be over cutoff. We therefore find it necessary also to remove correlations between words occurring over 100 times (half our document collection size in this particular test).

The correlations remaining after these cutoffs are applied represent non-random word co-occurrences. This does not necessarily imply that the words are related semantically. Co-occurrences may result from quirks of an author's style, or from peculiarities of word usage within document collections, as well as from actual semantic similarity. Since it has been suggested that word associations can be used to construct thesauruses, it is important to know whether word-word pairs produced by an association process reflect semantic meanings.

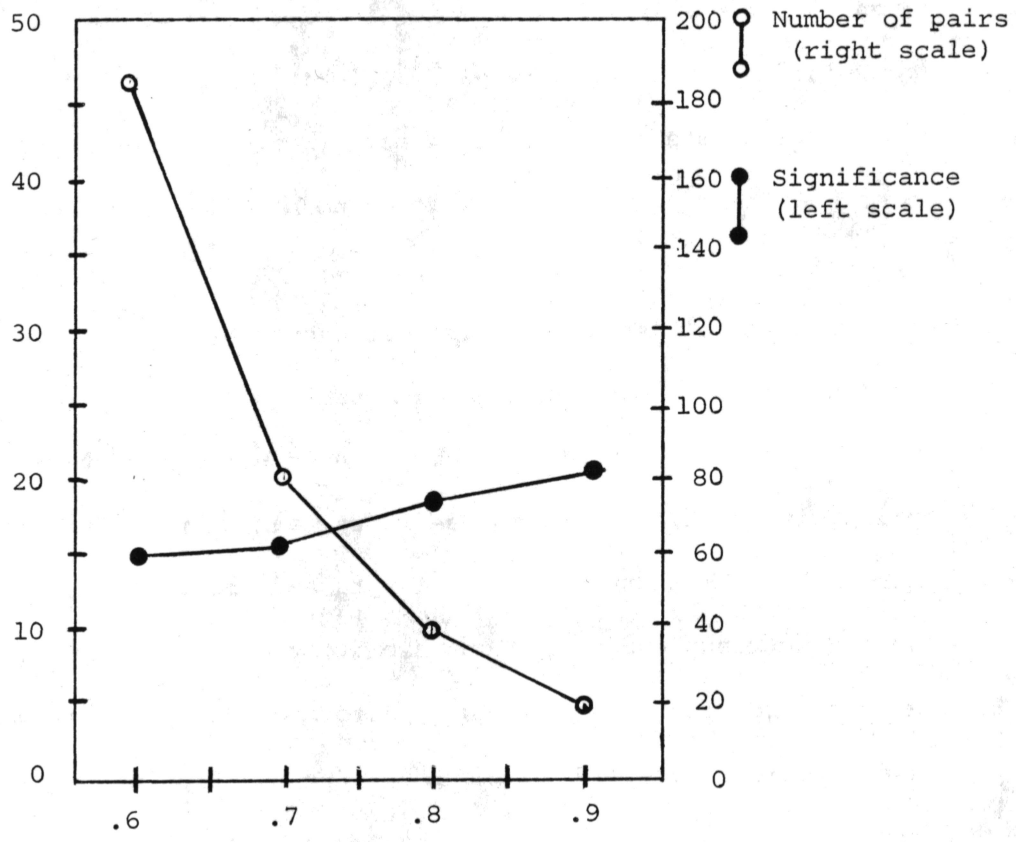
To investigate this problem, a list of word-word associations from a collection in aerodynamics (the Cranfield collection mentioned earlier) was prepared and analyzed for significance, the cutoff used being 0.60. Each word pair was examined and judged either as significant or non-significant. Significant pairs are those pairs which seem to be composed of semantically related words. The words are judged to be semantically related if they would normally be used together in discussions of the same topic, considering the most common technical definitions of the words to be their meaning. For example, "per" and "cent" is considered a significant pair; so is "atmosphere" and "satellite". On the other hand, "leading" and "edge" is judged non-significant (by this standard) as is "km" and "200", or "machine" and "evaluating". There is a great deal of subjectivity in such decisions; however, all judgments were made by the author and are therefore reasonably consistent.

All the word pairs were then classified by frequency of components and by correlation. The resulting table was then examined to see if any combination of parameters yields a particularly high ratio of significant to non-significant correlations. Overall, only 16.2% of all correlations are judged significant. Fig. 2 and Table 2 show the variation of significance with correlation level. There is a small increase in the fraction of pairs judged significant at higher correlation, but the number of correlations above cutoff decreases so rapidly as the cutoff is decreased that this is not a practical way of improving the quality of information produced by the association scheme. Even at a cutoff of 0.9 (which means that if two words occur five times each, all five would have to be co-occurrences, or if each occurs ten times, nine would have to be co-occurrences)

Cutoff	Significance	No. of Pairs
.6-.7	15.6%	185
.7-.8	16.0%	75
.8-.9	17.3%	52
.9-1	20.0%	20
all	16.2%	332

Effect of Correlation Level on Significance

Table 2



Variation of Significance with Cutoff

Fig. 2

only 20% of the correlations were judged significant. High cutoffs should not be used in the word-word association process if the aim is to recover a sizable number of significant pairs.

The classification by word frequency also yields no particularly superior choice of options. Table 3 shows the variation of significance with the frequencies of the words in the associated pair. High frequency words, as might be expected, show somewhat more reliable relationships; but the amount of statistical scatter in this corner of the table (since the number of high-frequency words is so small) renders the numbers doubtful. In any event, even the best numbers (e.g. correlations of words above 20 occurrences, based on 8 pairs) are relatively poor; only 37% significant relations. There is, in short, no choice of frequency or correlation cutoff which yields reliably significant pairs. Examination of more complete tables showing both frequency and correlation dependence of significant pairs also discloses no particularly good combination. It is believed, then, that for a collection of the size used (40,000 words) the statistical association process cannot be used to yield reliable indications of generalized word meanings.

Confirmation of this comes by comparison of word association pairs with dictionaries, phrase lists, and hierarchies. The IRE collection, for which a thesaurus of 700 concepts and about 3000 stems, a phrase list of 400 entries, and a complete hierarchy exist, is used for this test. A word-word association run was performed and the pairs were checked against all of these dictionaries. The association process identifies only one pair which is considered synonymous by the thesaurus; no pairs are

Frequency of words	3	4	5-6	7-8	9-10	11-19	20+	all
3	19%	10%	11%	13%	16%	14%	15%	14%
4		0%	9%	0%	40%	18%	20%	12%
5-6			12%	33%	16%	22%	19%	21%
7-8				0%	0%	20%	0%	5%
9-10					0%	43%	25%	25%
11-19						50%	29%	33%
20+							37%	37%

Dependence of Word-Pair Significance On Word Frequency

Table 3

Note: the 'all' column shows the percentage of significant correlations among all pairs which have the frequency of their lower frequency word indicated by the row frequency; not the percentage among all pairs with a word of the indicated frequency either as maximum or minimum frequency component.

included as phrases in the phrase dictionary; and only two pairs represent words which are directly related through the hierarchy.

An attempt to see what would happen if larger collections were used was also fruitless. A collection of 110,000 words (the ADI collection) yielded 19.7% significant correlations, ranging from about 10% in the lower frequency ranges to 50% in the higher frequency ranges. Because of the extreme length of documents in this collection, these results are not properly comparable with those from collections of abstracts, and further work on much larger collections is needed to determine whether reliable word relationships can be obtained from longer collections.

Since few of the word associations represent obvious semantic relationships, it may well be asked what causes the associations. The answer seems to be that they represent relations of "local" semantic meanings peculiar to this collection of documents. That is, the meaning of a word in one particular document collection may differ widely from the normal meaning of the word. When the meanings of the word <u>in the collection</u> are considered, it is found that about 73.1% of the pairs are significant, in this "local" sense. For example, consider the associated pair "scheme" and "machine". This was rated non-significant, since the words in their normal technical meanings are not related. However, it is found in examining the ten occurrences of "machine" that all ten imply "digital computing machine"; although the collection discusses compressors, engines, etc. none of these are referred to as "machines". The "local meaning" of "machine" is therefore "computer". Similarly, the major local meaning of "scheme" turns out to be "algorithm" (i.e. not just any kind of plan, but a plan for a digital computer program). Clearly, "algorithm" and "computer" is a significant pair.

To determine the fraction of "significant" pairs on the local basis, the list of pairs was rechecked for significance and each word looked up in a concordance of the text to determine its local meaning. The results are shown as a function of frequency in Table 4, and as a function of cutoff in Table 5. Nearly three-quarters of the pairs are now meaningful. The remaining pairs which are not composed of related words are generally stylistic quirks. For example, the word "addition" was used only as part of the phrase "in addition", which appeared only in a few abstracts. The word "addition" was thus associated with the other words in these abstracts even though it had no significant meaning in this collection.

More often, however, non-significant pairs are derived simply by accidental preferences of the author or one or more abstracts for certain words. If one abstract contains many instances of one word, a few instances of another word in that same abstract may appear to be a major amount of overlap to the association routine. Non-significant pairs, however, represent only a small amount of the total number of pairs of words of high frequency when local meanings are taken into account.

The majority of associations represent such "locally" related words. Overall, about three-quarters of the associations consist of related words; and 80% of these are related only because one of the words has a peculiar meaning in this collection. Fig. 3 shows additional examples of these local meanings. As a result of this peculiarity, the association process is not directly useful for determining word pairs

Word_1	Word_2	Local Usage_1	Local Usage_2
km	density	altitude in earth's atmosphere	density of air in atmosphere
km	200	altitude in earth's atmosphere	height in km of atmospheric phenomena
cold	turbo-jet	temperature of exhaust from jet engine	turbo-jet
models	wire	models	material of which models are built
eccentricity	contracts	ellipticity of satellite orbit	contraction of satellite orbit
leading	edge	leading edge of shock tube front	front in shock tube

Associated Pairs with Local Meanings

Fig. 3

Frequency of words	3	4	5-6	7-8	9-10	11-19	20+	all
3	61%	55%	59%	73%	83%	77%	77%	68%
4		67%	65%	80%	80%	82%	70%	72%
5-6			100%	80%	79%	78%	78%	81%
7-8				67%	100%	100%	64%	78%
9-10					50%	100%	100%	94%
11-19						50%	67%	63%
20+							88%	88%

Significant Associations Using Local Pairs

Table 4

Cutoff	Significant	No. Pairs
.6-.7	71%	185
.7-.8	79%	75
.8-.9	76%	52
.9-1.0	95%	20
all→	71%	332

Effect of Correlation Level on Local Significance

Table 5

that should be connected in a thesaurus. It can be used, however, to point to word relations not normally apparent, and thus it serves as an aid to dictionary constructors who are working with a known collection.

It should be noted again that these experiments were run on a collection of 40,000 words. It may well be that in larger collections, the apparent meanings of words approximate their common meanings more closely. This point will be the subject of future investigation, but the presence of apparently meaningless correlations has already been noted by workers with much larger collections. [1]

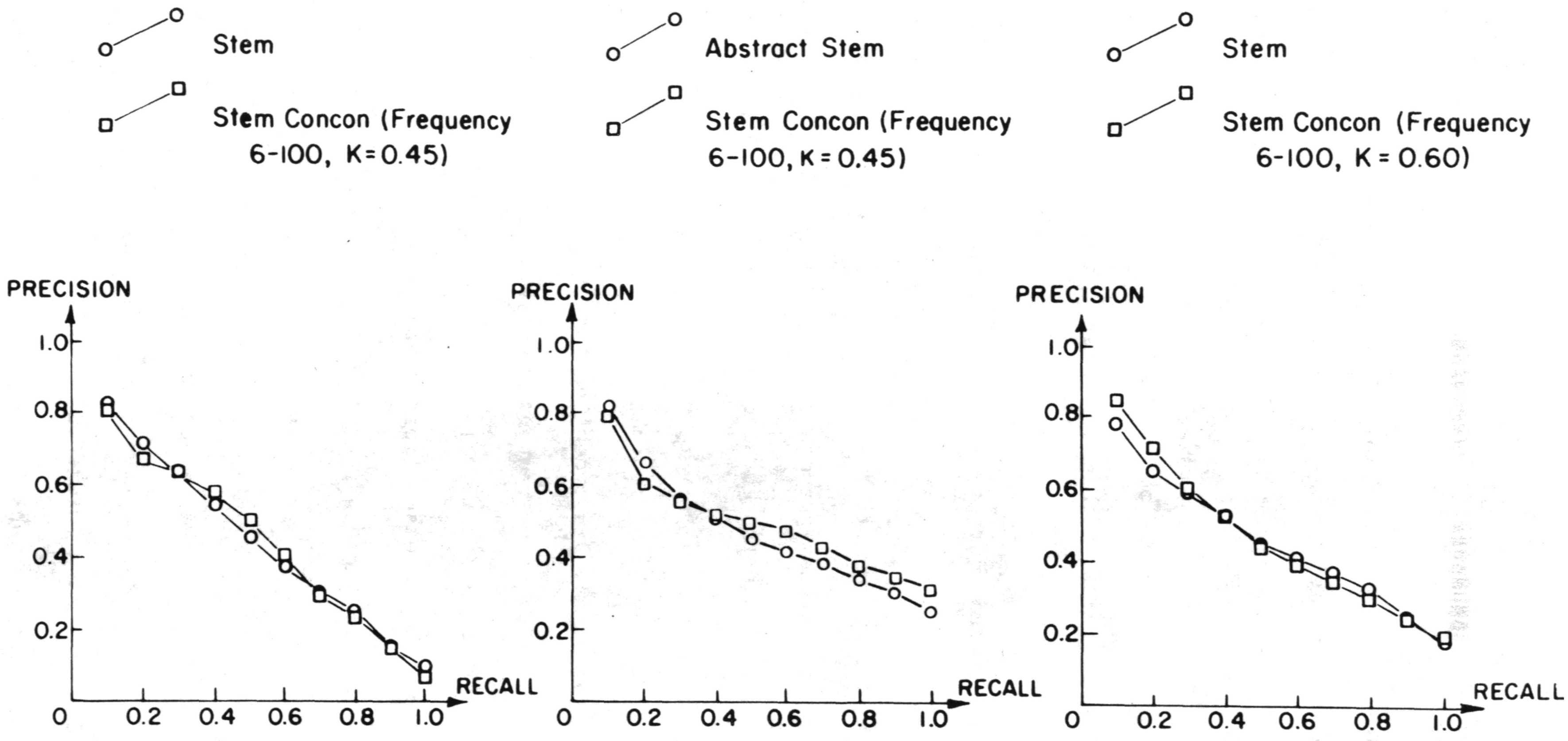
The properties of second-order associations were also investigated. These are word pairs, which need not co-occur in any documents, but must have common first-order associations. Almost all second-order associations, however, were also found to be first-order associated terms. They generally arise from large blocks of words, all of which were used to discuss some subject, and all of which were first-order associations of each other. For example, the set of words "height", "atmosphere", "density", "km", etc. are all used in a set of documents about the measurement of the density of the upper atmosphere. They were all identified as first-order association, and all became second-order associations. Stylistic quirks were not eliminated by the repetition of the correlation process; and the total number of associations was greatly diminished by a factor of 8-10. Second-order associations did not produce useful synonyms; even the one or two useful synonyms in the first-order associations (e.g. "error", as in "error function", and "erfc", its abbreviation) tended to disappear in second-order, as did most other associations. The use of second-order associations appears to offer no advantages over first-order associations, and loses a great amount of material.

We can therefore conclude that the use of associative procedures for the determination of word meaning in a general sense is not advisable with moderate sized collections of text, since the vast majority of the associations produced reflect specific local meanings of words. No choice of word frequencies or associative procedures appears to offer a way around this difficulty.

4. Retrieval Experiments

Although the association process is not suited for investigations of absolute word meanings, it is nevertheless useful in retrieval systems. Fig. 4 shows a comparison between word-word association retrieval runs and straight word stem matching for three collections. It is seen that for two of the collections (the ADI and IRE collections) the improvement offered by associative strategies is only over small ranges and of doubtful significance. For the Cranfield collection, the associative strategy shows a definite superiority.

The purpose of the associative method, originally, is to produce word relations missed by the stem matching procedure, and thus to take the place of a synonym list or thesaurus. It would be expected that such a procedure would be a recall-oriented device. However, this is not quite what happens. Associative procedures improve performance in two distinctly different ways. First, they do occasionally retrieve a document that is missed in stem matching, by introducing new word relations which provide some request-document overlap. More often, however, precision is improved



IRE-3, 34 Requests

Cranfield-I, Abstracts, 42 Requests

ADI, Full Text, 35 Requests

Comparison of word stem dictionary with addition of Statistical Word-Word Association (Stem Concon).

Fig. 4

by promoting documents which were already retrieved but at a moderately low level. This is done by increasing the weight of significant words which previously matched in the request and document by adding associated words to both. This increase in the weight of the significant words (by addition of words which co-occur with them) improves performance. This second effect is in fact responsible for most of the improvement shown by the association process.

The precision effect might seem to correspond to the role of the list of non-significant words in the thesaurus method, just as the recall effect corresponds to the synonym list. In practice, however, the non-significant words needed for this purpose are often missed in thesaurus construction. For example, the thesaurus constructor may easily fail to recognize the uselessness of "addition", "hand", "order", and "example", if he does not know that they have occurred only in the combinations "in addition", "on the other hand", "in order to" and "for example" in the particular collection. Also, high frequency words, even if they retain their semantic sense, are often of no value for retrieval because they occur so often as to provide no discrimination between documents. Unless the thesaurus is made with the aid of a complete concordance, such errors are quite likely to occur.

The two effects of the associative process can be seen in Table 6. This shows the changes in rank position of relevant documents from the word stem matching process to the associative retrieval run, using a term frequency range of 6 to 50 and a cutoff of 0.45. The number of documents which change from each range of rank positions in the associative run is shown in the main block of the table and the net change in each rank group is shown at the bottom. It is observed that the associative process removes relevant documents from ranks 10-29, but not 30-99. It also removes documents (but many fewer) from the very low rank positions. The documents in ranks 10-29 which are promoted up to ranks 1-9 are generally those which have had their significant terms upweighted by the process indicated above. The majority of documents promoted from ranks 100-200 had no significant matching terms before associations were added. Of 6 relevant documents which move up over 100 rank positions, all were improved by the recall effect. But in the 10-29 groups, which lost a total of 38 relevant documents, 25 going up and 15 down, nearly every case of improvement is due to the precision effect. This represents a significant improvement in the performance of the relevant documents in this range. In fact, the largest change of ranks in the entire table is the promotion of 16 relevant documents from the 10-19 range to the 1-9 range. In the 100-200 range, a total of 14 documents were promoted out of this range, while 11 were dropped down into it. But it should be noted that if the range 20-99 is considered, as many documents (10) are dropped from these ranges to the 100-200 range as are promoted to them. The net loss of documents from the 100-200 range is due entirely to the 4 promotions to rank positions 1-20, all caused by the recall effect.

Another way of describing this effect is to note, for each rank position range, the number of relevant documents promoted as against the number of relevant documents demoted. This is shown in Table 7. Clearly, the system operates well near the 20-29 range; and then again at the very bottom. The last figure in the table is somewhat inflated, since if a document is already near the bottom, it is difficult to demote if further.

		Rank in Association Retrieval Process						
		1-9	10-19	20-29	30-49	50-99	100-200	total
Rank in Word Stem Process	1-9	67	12	0	4	3	1	87
	10-19	16	7	4	5	4	0	36
	20-29	7	2	2	2	1	3	17
	30-49	2	0	1	3	5	2	13
	50-99	2	2	1	1	11	5	22
	100-200	2	2	1	2	7	9	23
	total	96	25	9	17	31	20	
	net change	+9	-11	-8	+4	+9	-3	

Changes in Rank Positions of Relevant Documents for Two Analysis Methods

Table 6

Rank Positions	Promoted	Demoted	Ratio
1-5	16	35	.46
6-9	9	13	.69
10-19	18	16	1.13
20-29	10	7	1.43
30-39	2	2	1.00
40-49	3	7	.43
50-99	10	11	.91
100-199	17	3	5.67

Promotions and Demotions as Function of Rank Positions

Table 7

The mechanism of precision improvement, as previously stated, is to re-inforce the apparent weight of significant terms by adding their associated terms. This process works because the significant terms generally have more associated pairs than the non-significant terms. It may be asked why this should be so. This feature of the associations derives from the greater concentration of the significant terms in the abstracts. The non-significant terms are generally widely spread among the abstracts, so that it is difficult for any term to match their occurrences. The significant terms are clustered in a few abstracts, and another term can match them easily, since only one or two co-occurrences of terms which occur several times in each document in which they appear is necessary to produce a correlation above cutoff. That is, if terms occur once in each of ten documents, they must occur in six common documents to correlate at a 0.6 level; but if they occur three times in each of three documents, they need only co-occur in two documents to correlate at a level of .67. The tendency of significant terms to bunch up is shown in Table 8 which shows the distribution of occurrences of the ten words occurring fifteen times. It is seen that no non-significant term occurs in fewer than twelve documents; none occurs more than three times in a document; and their average ratio of number of occurrences per document is only 1.1. The significant terms never occur in more than ten documents; every one appears at least three times in some document; and they average 1.8 occurrences per document.

An example of the effect of this is shown by query Q116. This query contains twelve words in the word stem matching system, of which the key word is "dissociated". "Dissociated" was outweighted in the search by such high-frequency words as "wind", "high", "pressure", etc.

Word	Significant?	No. of Docs.	Occurrences								Occ./doc.
			1	2	3	4	5	6	7	8	
airfoil	yes	8	3	3	2						1.9
affect	no	13	11	2							1.2
converge	yes	8	4	2	1	1					1.9
correct	no	14	13	1							1.0
km	yes	7	5	1						1	2.1
magnitude	no	14	13	1							1.0
plan	yes	10	7	2		1					1.5
practical	no	12	10	1	1						1.25
previous	no	14	13	1							1.0
vortex	yes	7	4	1		1	1				2.1

Distribution of Occurrences of Significant and Non-significant Words

Table 8

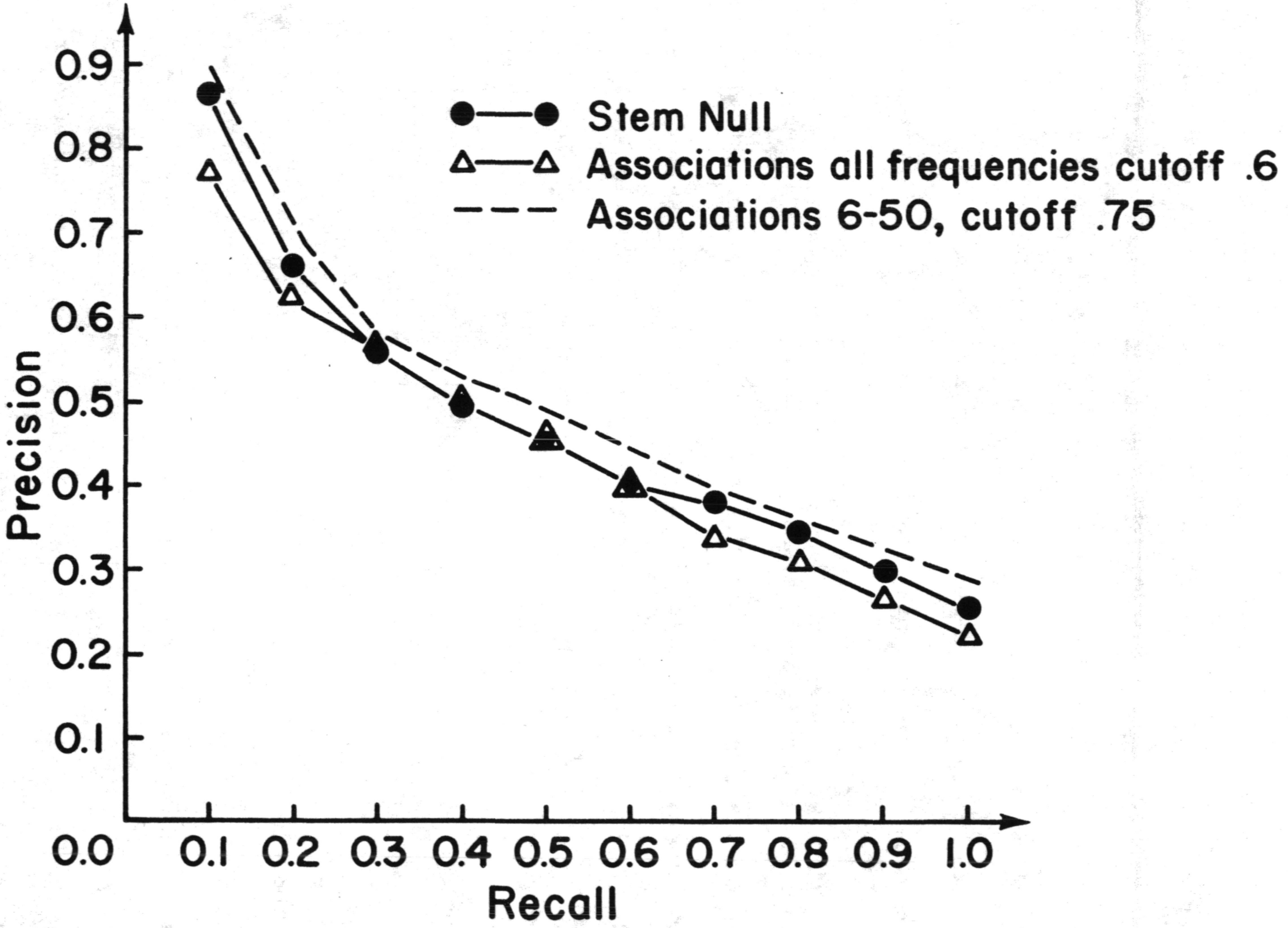
which are too frequent in this collection to be of much use as search terms. However, none of those words had any related pairs, while "dissociated" introduces eight new words. As a result, while "dissociated" represented only 8% of the original query, it and its associations represented 28% of the new query. The additional weight given to this important term (since all its associations are also introduced into any document which contains the word) causes three documents in rank positions 21, 23, and 27 to be promoted to positions 1, 6, and 7. Note that "dissociation" already appears in these documents before expansion; but it is not emphasized enough.

Recall-effect improvement (introducing new terms missed in the original search) is illustrated by a question in the ADI collection, QB2, on the "testing of automatic information systems." This fails to match one relevant document which deals with the "evaluation of documentation techniques". The association procedure connects "automated" in the query with "experiment" and "reduce"; "reduce" in turn is related to "documentation". This provides enough overlap to raise the document from 77th place in the rank list of retrieved documents to 9th. It should be noted that the useful relations are locally significant pairs (e.g. "automated" and "experimented"; "experiment" and "test" are not associated).

An example from the Cranfield collection is query 226, whose key term is "Navier-Stokes" (equation). Document 08C does not contain this word, but it was introduced by the association procedure from the word "steady". The word "numerical" was introduced into both query and document from "Navier-Stokes" and "steady", respectively. Again, note the local significance of these pairs; the thesaurus does not connect "Navier-Stokes" with any of these terms. As a result of these associations, this relevant document is promoted from rank position 143 to rank position 4.

The results of retrieval experiments can be used to determine the best set of parameters for the association process. The conclusions agree well with those deduced from the examination of the pairs in part 3. It is noted there, for example, that words that are either very frequent or very rare tend to have non-significant associations. The fraction of meaningful correlations can also be increased by raising the cutoff. The effect of this on retrieval is shown in Fig. 5, where recall-precision curves for the stem dictionary directly — without any associations added — and for two different association strategies, are compared. When all words, of whatever frequency, are used in the association process, the resulting curve is usually inferior to the normal word matching run. But when the frequencies of words employed in the association process are restricted to the range 6-50, and the cutoff is raised, the resulting recall-precision curve is everywhere superior to the stem curve.

It is also noted in part 3 that words occurring only three or four times have fewer significant occurrences than words of six or more occurrences. The effect on retrieval of variations in the frequencies of words used in the association process is shown in greater detail in Table 9. For both recall and precision purposes, the optimum frequency range appears to be 6-50, although the differences in performance are small. Examination of the recall-precision curves of Figs. 6 and 7 shows the frequently crossing curves, and thus the insensitivity to



Recall-Precision Curves for Association Runs - Cranfield Collection

Limitation of Association Process

Fig. 5

Minimum Frequency	Maximum Frequency			
	25	50	100	none
none				.2991
3	.2840	.3051	.3126	.3070
6	.2933	.3162	.3148	.2887
10	.2838	.3123	.2970	.2561

a) Rank Recall for Different Frequency Ranges

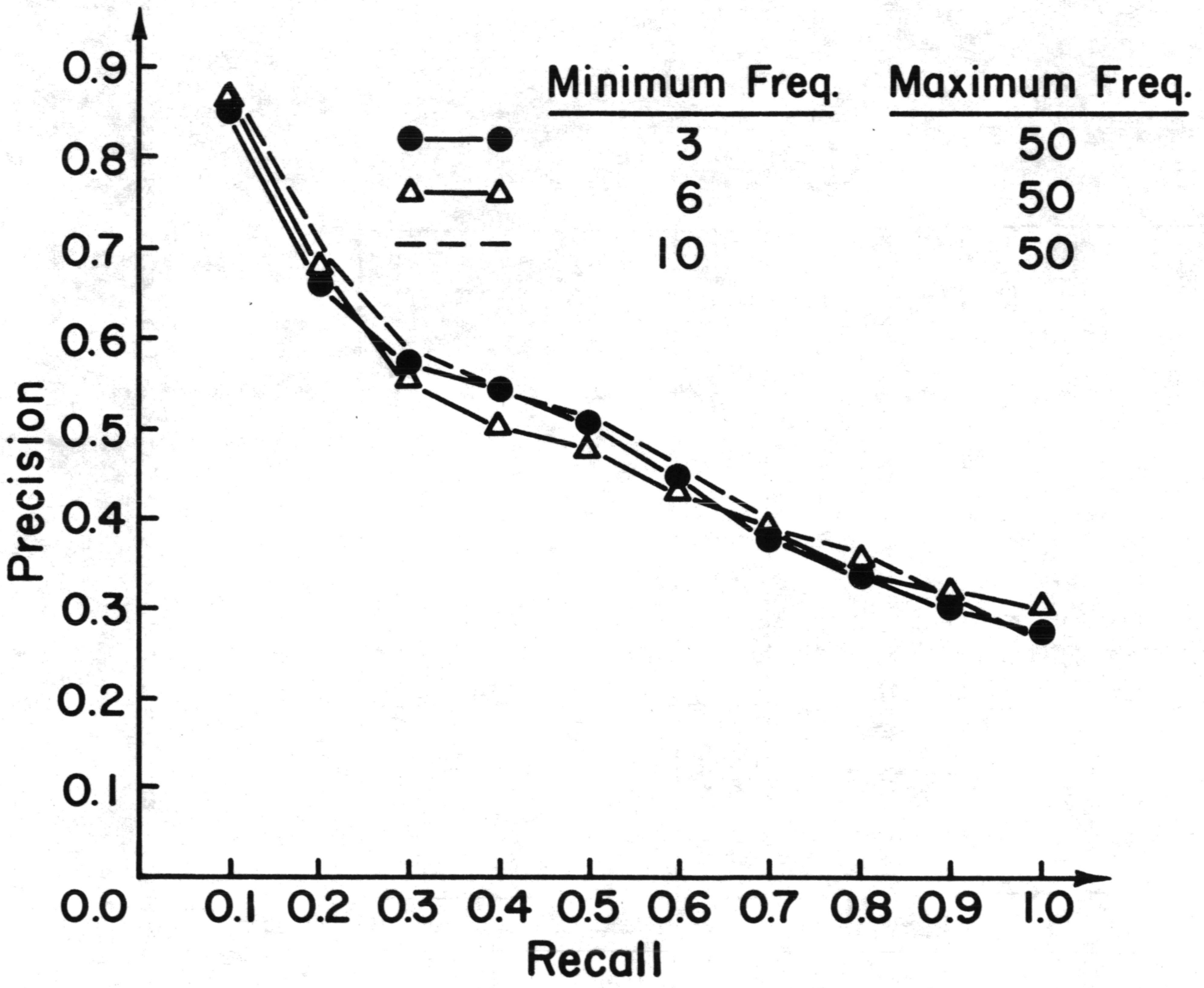
Minimum Frequency	Maximum Frequency			
	25	50	100	none
none				.4642
3	.4684	.4823	.4756	.4647
6	.4711	.4888	.4738	.4453
10	.4488	.4737	.4567	.4179

b) Log Precision for Frequency Ranges

Rank-Recall and Log Precision for Various Frequency Ranges

Table 9

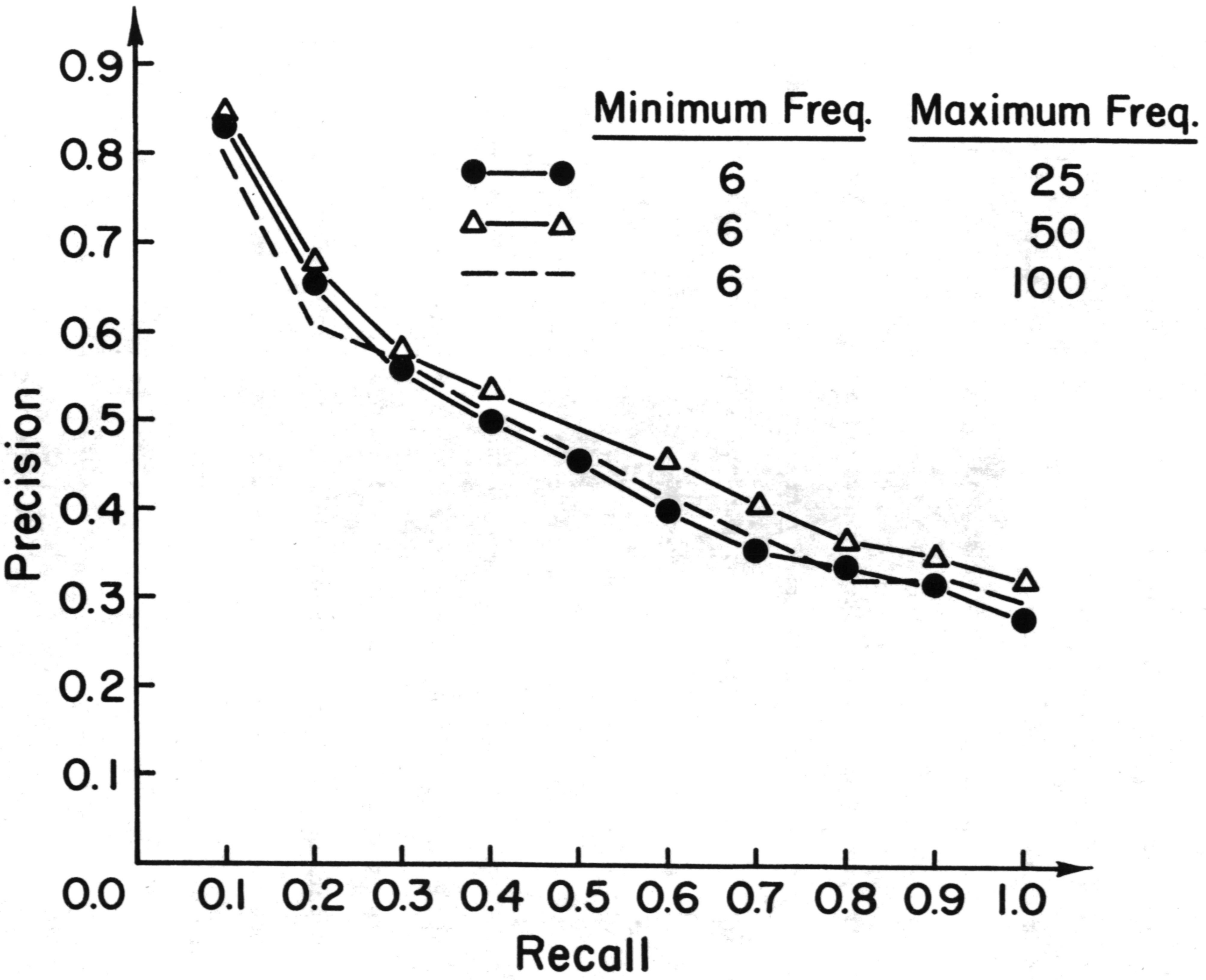
All runs with cosine correlation, cutoff 0.6, expansion weight 1.0



Recall-Precision Curves for Association Runs - Cranfield Collection

Variation of Minimum Frequency

Fig. 6



Recall-Precision Curves for Association Runs - Cranfield Collection

Variation of Maximum Frequency

Fig. 7

these parameters. The majority of the improvements obtained by restricting frequency of words processed is obtained by removing the associations involving words of frequency 1 and 2.

A comparison of the two correlation algorithms (cosine and overlap) is shown in Fig. 8. These curves also cross several times, and neither correlation coefficient can be called superior. The cutoffs used in the two methods are chosen to roughly equalize the number of associated pairs. As the cosine algorithm was designed primarily to handle the request-document correlation problem, in which the vectors are of widely different length (which is not so often the case in the present problem, since the extremely rare and the extremely frequent concepts are omitted), it is not surprising that the algorithms perform similarly. Since neither correlation coefficient shows a distinct advantage, the cosine correlation is used in all other retrieval runs described in this section.

The effect of varying the cutoff used in the association process is shown in Table 10 and Fig. 8. Again, the curves cross, with the lowest cutoff being superior at high recall and the highest cutoff being superior at high precision. As a high cutoff produces the fewest but most reliable associated pairs, it is expected to be preferable for precision purposes, whereas a low cutoff produces the largest number of significant pairs and therefore has an advantage if maximum recall is demanded. The cutoff of 0.9, however, is so high that such an association process is almost indistinguisable from the word stem run; and the cutoff of 0.3 is so low as to introduce large numbers of non-significant pairs. The useful range of cutoffs seems therefore to be 0.45-0.75, roughly, for the cosine correlation.

Table 11 shows the effects of varying the relative weight of the associations (a weighting of 1 renders a word introduced into a document

Cutoff	Rank Recall	Log Precision
.30	.3247	.4770
.45	.3409	.4982
.60	.3162	.4888
.75	.3116	.4932
.90	.2933	.4608

All runs cosine correlation, frequency range 6-50

Effect of Varying Cutoff

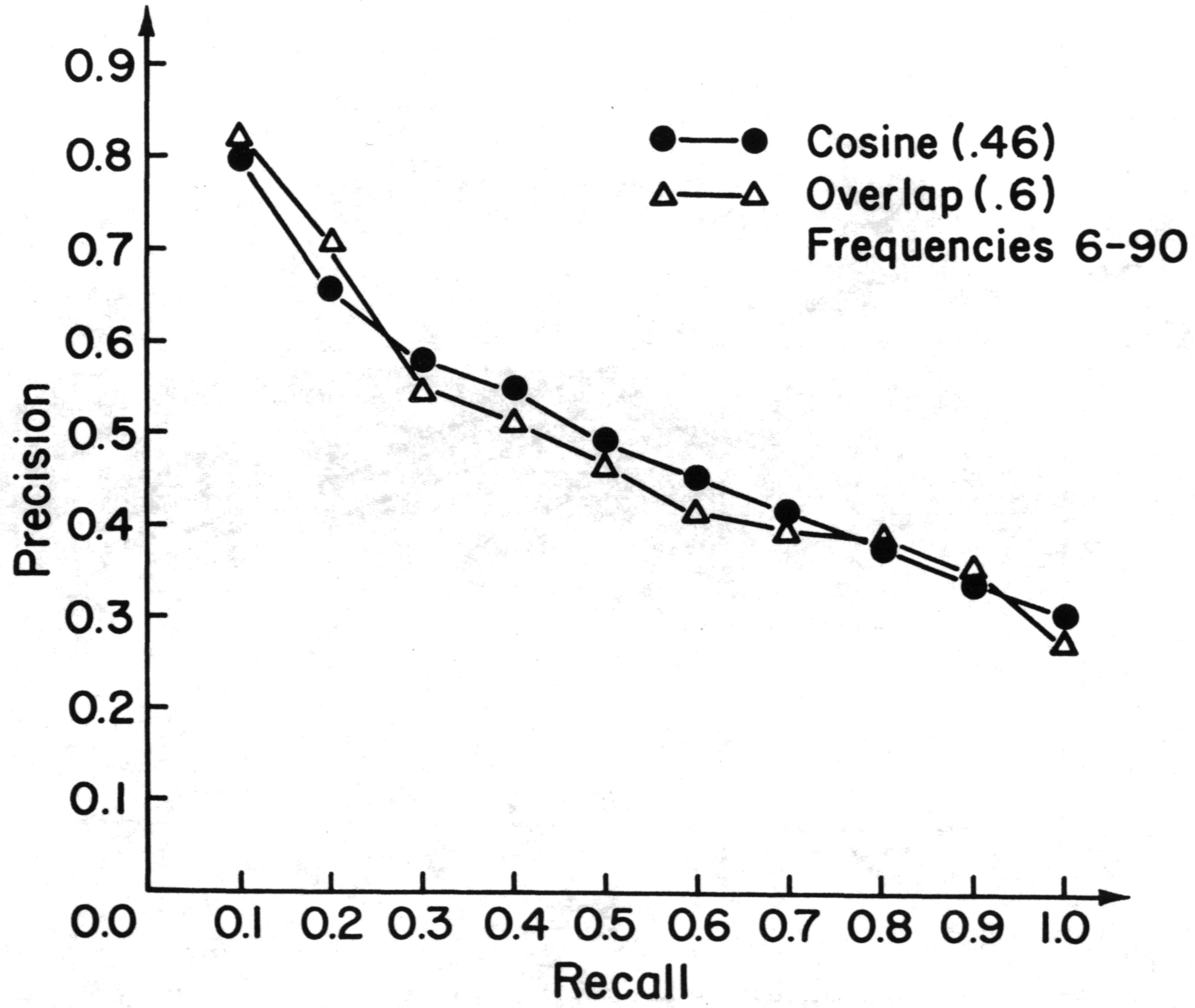
Table 10

Weight	Rank Recall	Log Precision
.125	.3009	.4668
.25	.3224	.4991
.5	.3239	.4952
1.0	.3162	.4888
2.0	.3007	.4712

All runs 6-50, cosine correlation, cutoff 0.6

Effect of Varying Weight of Associations

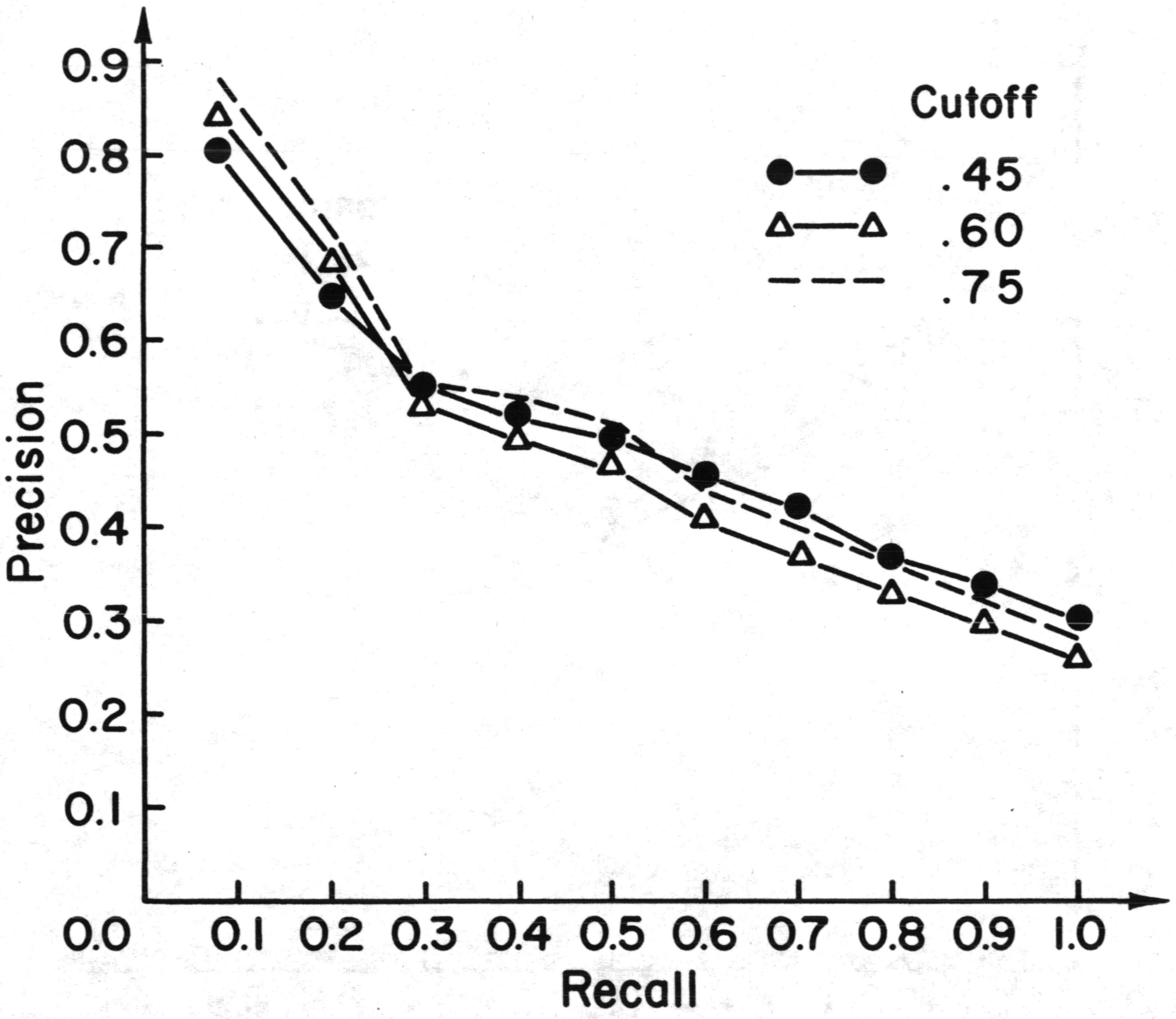
Table 11



Recall-Precision Curves for Association Runs - Cranfield Collection

Effect of Correlation Mode

Fig. 8



Recall-Precision Curves for Association Runs - Cranfield Collection

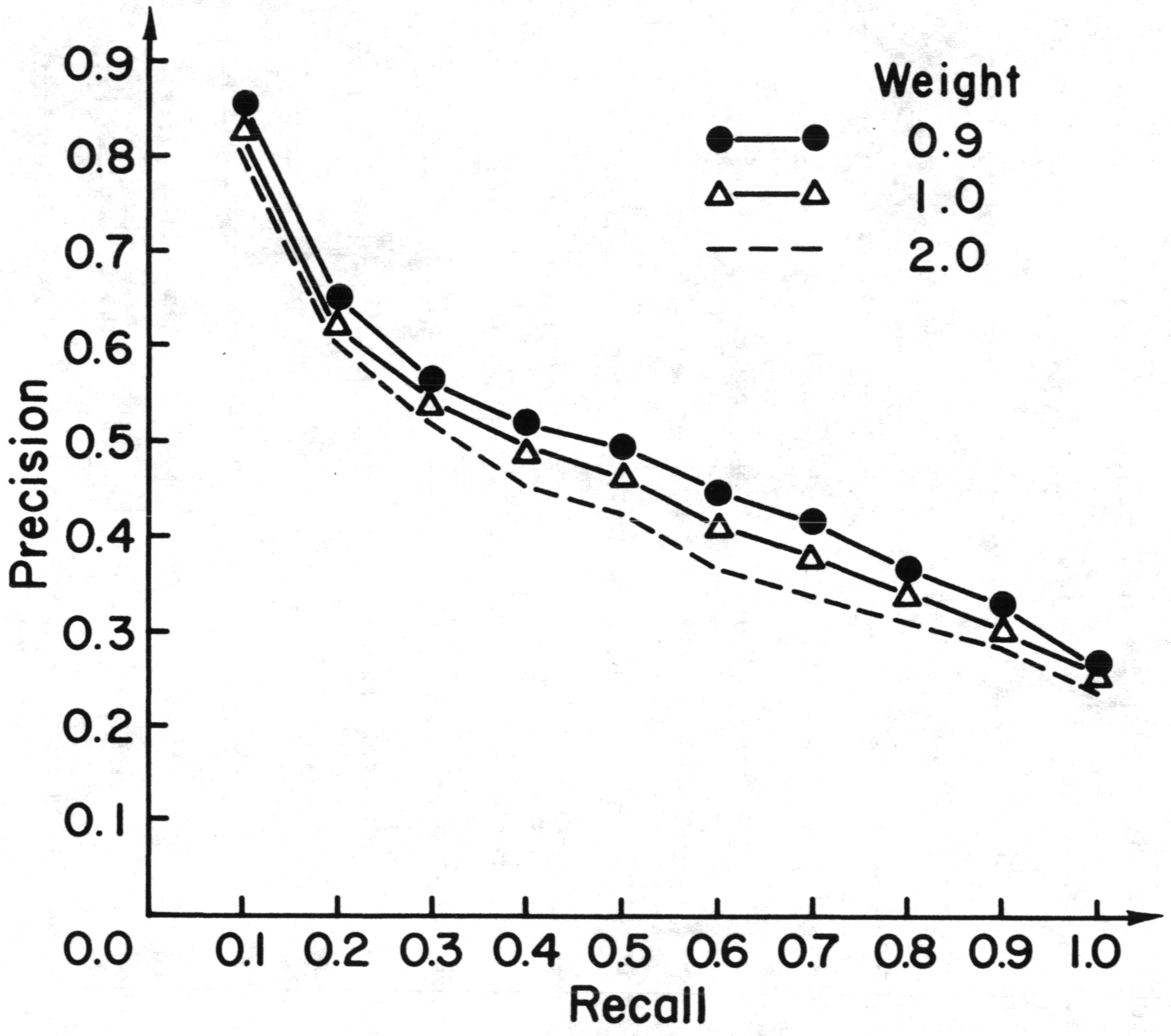
Variation of Cutoff

Fig. 9

vector by the association process equally important as a word in the original document). Weights somewhat below 1 are seen to be preferable, more so for precisions than for recall purposes. Fig. 10 indicates, in fact, that for high recall, weights above 0.5 do not cause as much loss in performance. To sum up, then, for high precision, one should have low weights and high cutoffs; for high recall, higher weights and lower cutoffs are desirable. Fig. 11 indicates recall-precision curves with high recall and high precision specifications; as expected, they cross.

It was also seen in part 3 that additional iteration of the association process is not useful in finding synonyms, and it is also not of great value in retrieval. Fig. 12 shows curves for 0, 1, and 2 iterations of the association procedure, with frequencies of 6-50 and a cutoff of 0.60. The first iteration curve is seen to be superior.

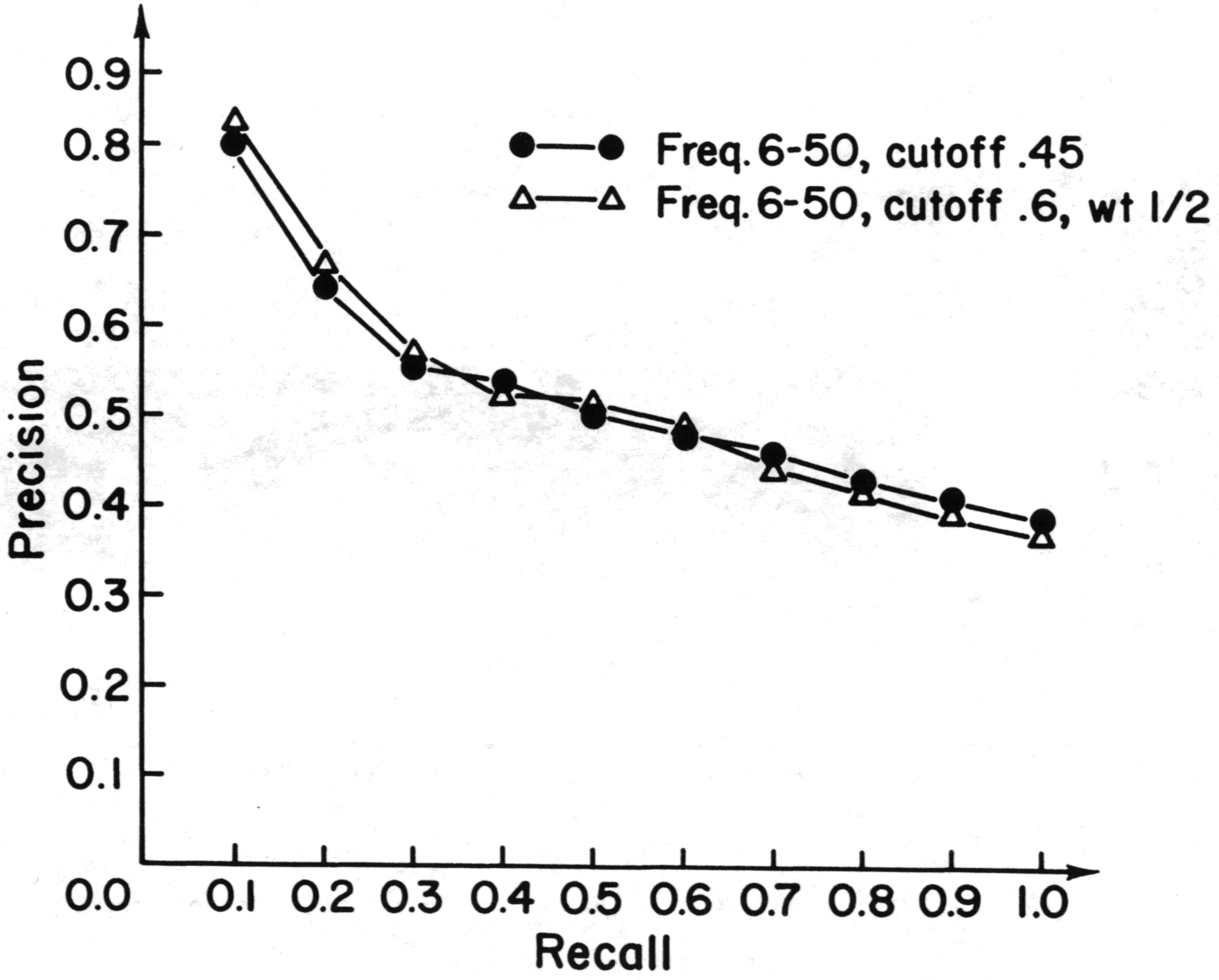
The performance differences shown by the various options in the association process are rather small. It is difficult, in particular, to choose a set of options to maximize either the precision effect or the recall effect over an entire set of requests. Nor does a fine adjustment of cutoff, frequency, or weight have a major effect on retreival performance. This is just what is expected from the analysis of the associated pairs, since no set of parameters produces an unusual number of significant pairs. In general, the use of associated pairs produces improvement in performance over most of the range compared with word, stem matching if words with very low and high frequencies are omitted. Procedures which decrease the number of associated pairs (restricting the frequency range used, raising the cutoff) or lower the weight of the



Recall-Precision Curves for Association Runs - Cranfield Collection

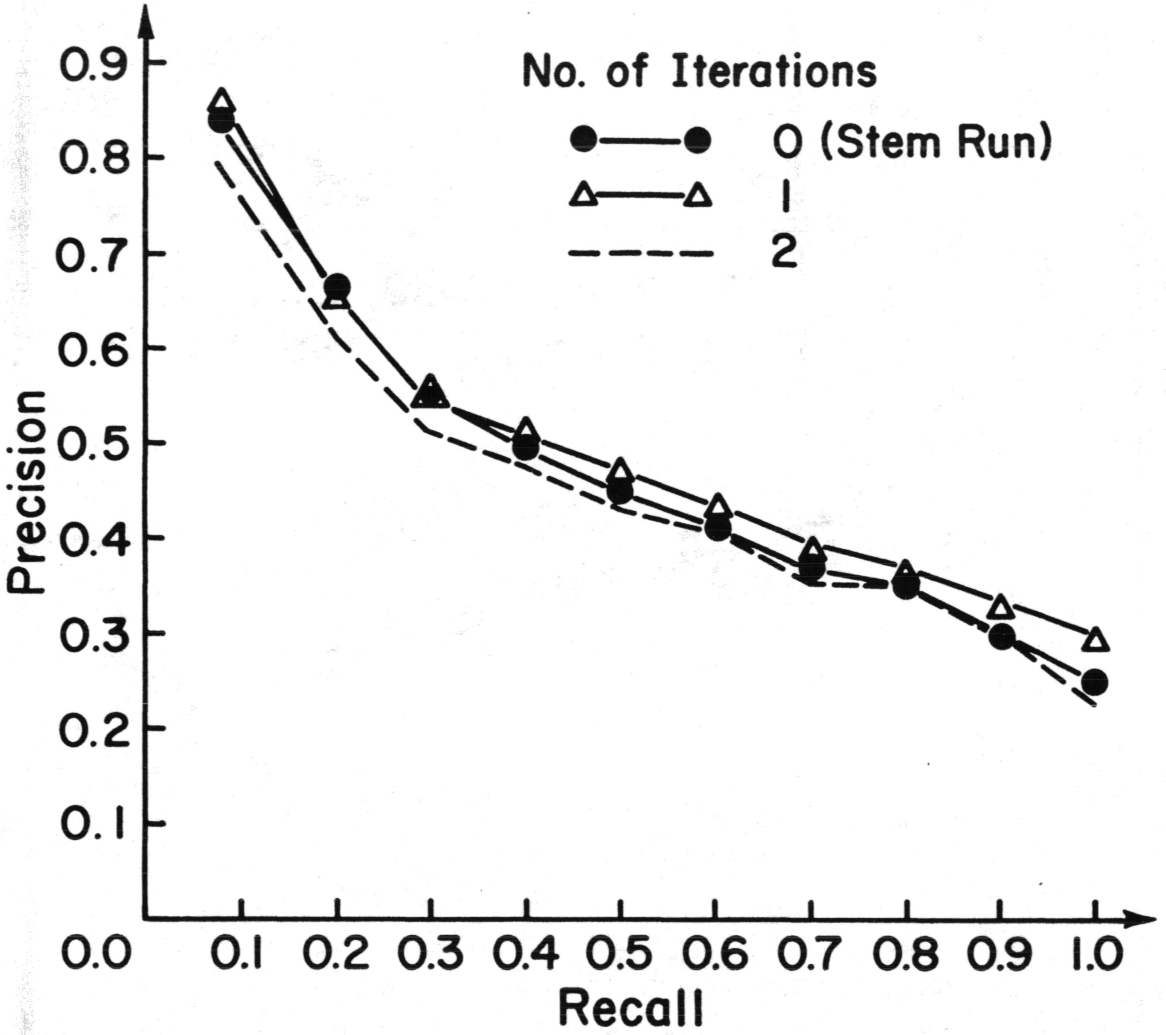
Variation of Weights

Fig. 10



Comparison of High Recall and High Precision Association Methods

Fig. 11



Recall-Precision Curves for Association Runs - Cranfield Collection

Effect of Iteration

Fig. 12

associations, are best for precision; procedures which increase the number of associated pairs or their weight are best for recall, but the effect is small.

It is often hypothesized that association procedures can simulate the operation of a thesaurus or other classical word normalization procedures. This can be tested on the Cranfield collection, since a thesaurus is available as well as a form of indexing. The indexing, although very detailed and exhaustive (averaging over 30 terms per abstract) is not carried through a rigorous term normalization, and the results with it may perhaps be unusual. It is felt, however, that in terms of overall performance, the exhaustivity and high quality of the indexing compensates for the lack of normalization, so that results should be roughly comparable.

As expected from the earlier discussions in this section, the association procedure operates in a unique and virtually independent way, simulating neither indexing nor thesaurus. Out of 42 requests, the thesaurus improves the performance of 24, the indexing improves the performance of 24, and the association procedure (with a frequency range of 6-100, cutoff of .45, weight of 1.0) improves 25. Yet only 12 requests are improved by all three methods (even on a random basis at least 8 would be improved by all 3). Table 12 shows two-by-two contingency tables for co-improvement of requests by thesaurus and association, and Table 13 shows two-by-two contingency tables for co-improvement of requests by indexing and association. None of the tables are significant, i.e. there is no co-variation of association results and thesaurus results of association results and indexing results. A set of recall-precision curves for thesaurus and associations is shown in Fig. 13, and for indexing and association in

	Request improved by association over null	Request not improved by association over null	
Request improved by thesaurus over null	15	9	24
Request not improved by thesaurus	10	8	18
	25	17	42

a) Improvements of Association Methods over Null and Thesaurus

	Request Greatly improved (rank recall up by .1) by association over null	Request not greatly improved by association over null	
Request greatly improved (rank recall up by .1) by thesaurus over null	5	3	8
Request not greatly improved by thesaurus over null	9	25	34
	14	28	42

b) Large Improvements of Association Methods over Null and Thesaurus

Contingency Table for Co-Variation of Thesaurus and Association Methods

Table 12

	Request improved by association over null	Request not improved by association	
Request improved by indexing over null	16	8	24
Request not improved by indexing over null	9	9	18
	25	17	42

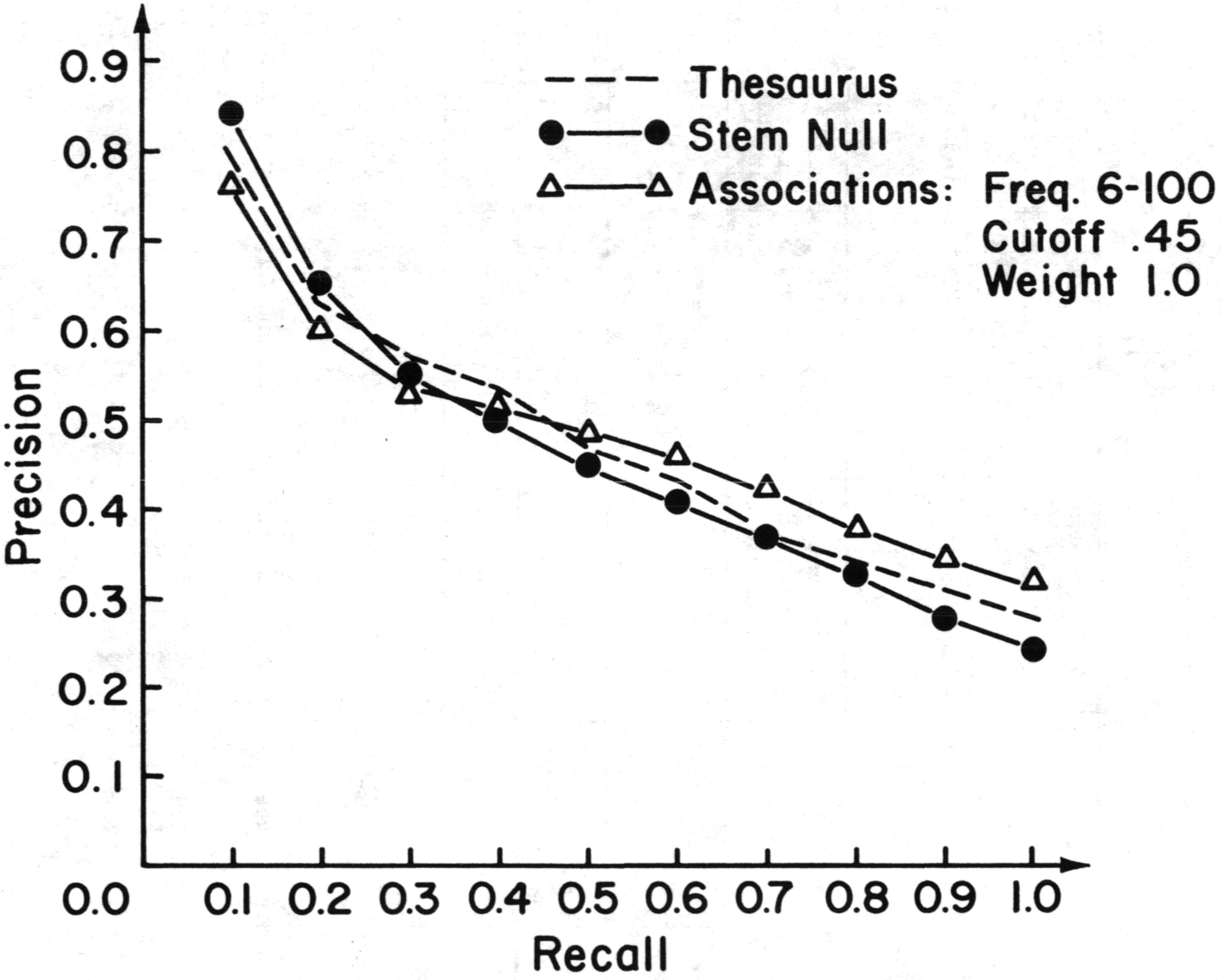
a) Improvements of Association Methods over Indexing

	Request greatly improved by association	Request not greatly improved by association	
Request greatly (rank recall up .1) improved by indexing	4	5	9
Request not greatly improved by indexing	10	23	33
	14	28	42

b) Large Improvements of Association Methods over Indexing

Contingency Tables for Co-Variation of Indexing and Association Methods

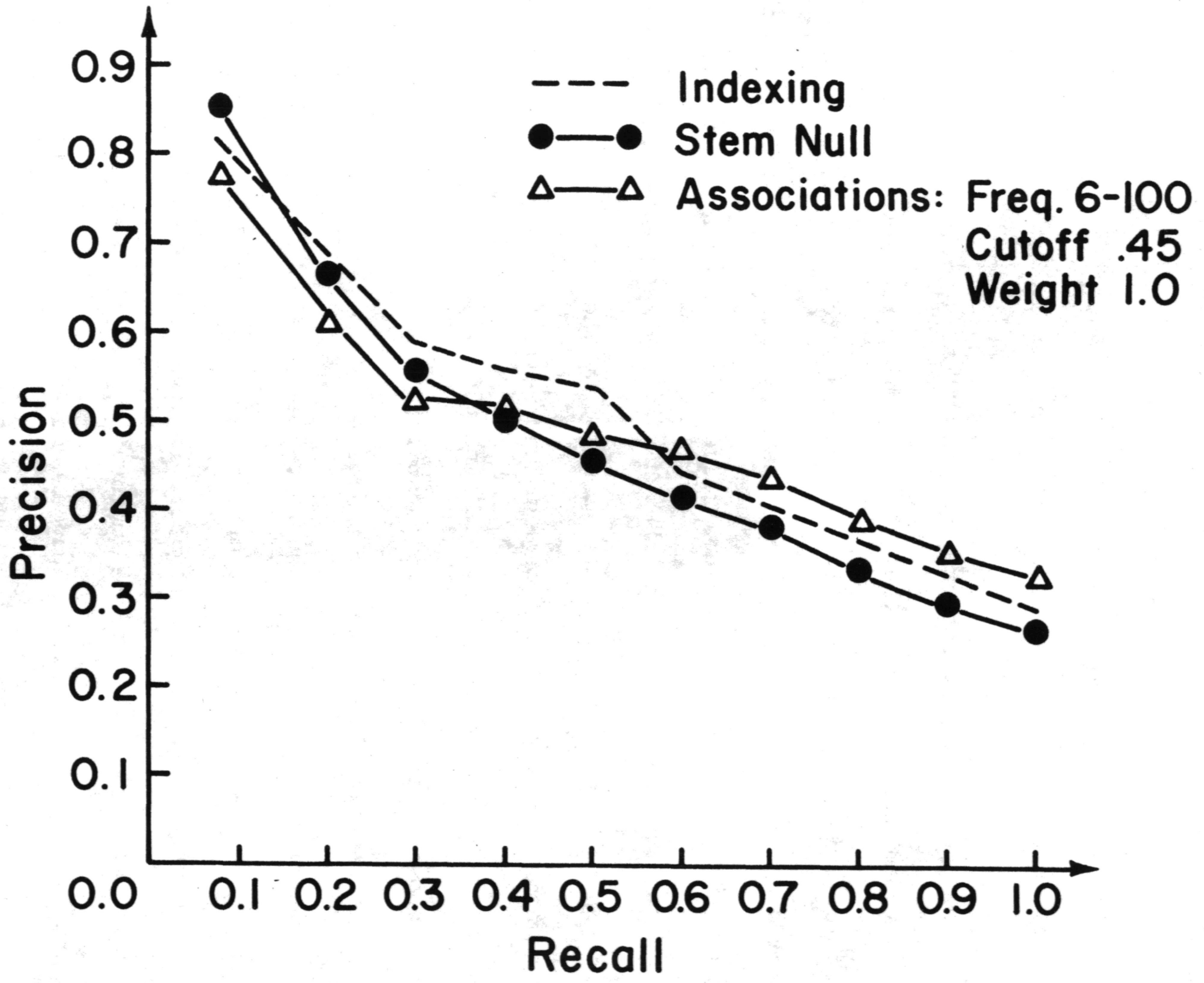
Table 13



Recall-Precision Curves - Cranfield Collection

Comparison of Association and Thesaurus

Fig. 13



Recall-Precision Cruves - Cranfield Collection

Comparison of Association and Indexing

Fig. 14

Fig. 14. The curves cross and the order of superiority depends on the operating range of the system. In short, the association process

a) does not detect the same word relationships as a thesaurus or an index,

b) does not use them in the same way, and

c) does not have the same effect on retrieval.

The operation of associative retrieval as a precision device can be illustrated also by noting that the mechanism proposed (promotion of relevant material near the top of the rank list) requires a moderately good performance to begin with. When the performance of associative retrieval methods are compared between requests with good and bad performance in the stem dictionary, it is seen that hardly any requests with bad performance in the stem dictionary are promoted by the association process. This is shown in Table 14. Although there are 20 requests with rank recall above .2, and 22 requests with rank recall below .2, 10 of the 14 requests showing performance improvement with the association process have a stem rank recall of greater than 0.2.

Finally, Table 15 shows the 25 requests which have a rank recall increase of at least .1 with some method, with a note of which methods improve the request. Only two requests improve by this magnitude on all three methods. More requests are improved by the associative process only, and not by the thesaurus, than are improved by both the thesaurus and the associative process. It seems fair to conclude that associative retrieval and the use of a thesaurus are essentially independent methods of improving request performance.

	Large improvement with association (rank recall up .1)	Not large improvement with association	
rank recall in null above .2	10	10	20
rank recall in null below .2	4	18	22
total	14	28	42

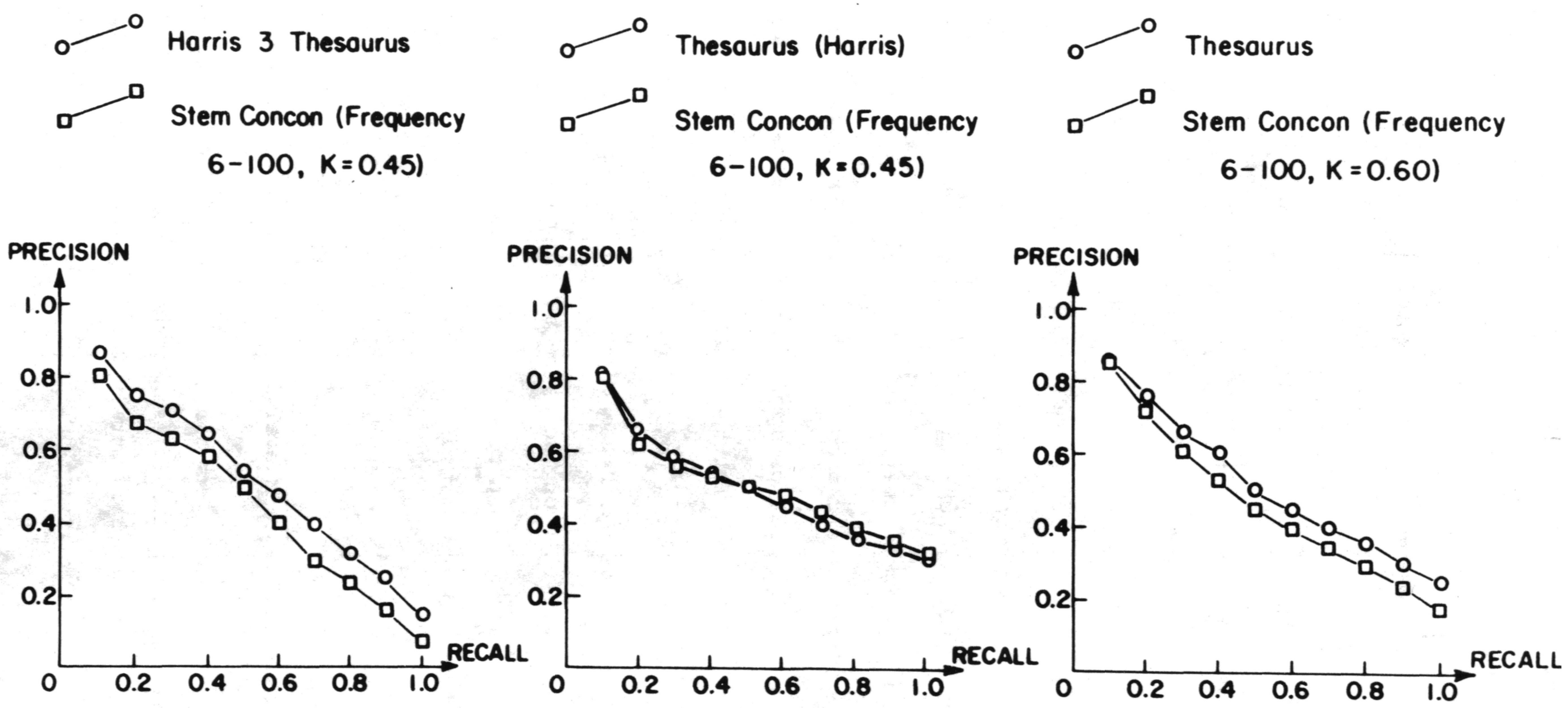
Contingency Table for Co-Variation of Association Improvement and Bad Performance in Stem Dictionary

Table 14

Query	Improved in Association	Improved in Thesaurus	Improved in Indexing
Q079			X
Q100	X		
Q116	X		
Q121	X	X	X
Q130		X	
Q136	X		
Q141	X	X	
Q145	X		X
Q146			X
Q148	X		
Q181		X	
Q189	X		
Q190	X	X	
Q226	X		
Q230			X
Q250	X		
Q268		X	X
Q269	X		X
Q273	X	X	
Q317	X	X	X
Q360			X

Requests with Large Improvement (rank recall up by 0.1)

Table 15



IRE-3, 34 Requests

Cranfield-I, Abstracts, 42 Requests

ADI, Full Text, 35 Requests

Comparison of Thesaurus Performance with Statistical

Word-Word Association (Stem Concon).

Fig. 15

Since the association procedure seems independent of a thesaurus procedure, one can ask which is the better method only in the sense that — considering the average collection and request — which method has a better probability of working well. Fig. 15 shows comparative recall-precision curves for thesauruses and association runs for three collections using the best association strategies. It is seen that for two of the collections, the thesaurus is definitely superior, and for the third collection (Cranfield), the difference in performance is non-significant. For the Cranfield collection, the thesaurus performs worse than for the other collections in general. It is believed that the reason for the poor performance of this thesaurus is that it was originally constructed for a different purpose, and thus is not properly optimized for the SMART programs. If this were to perform a little better, one would expect all three collections to show equivalent curves. Even with the performance curves shown in Fig. 6, however, it is clear that on the average, requests should be entrusted to a thesaurus rather than to an association scheme for maximum performance.

5. Conclusions

A survey of associative retrieval results indicates that

a) on small collections, associations are not for determining word meanings or relations, since the majority of the associated pairs depend on purely local meanings of the words and do not reflect their general meaning in the technical text;

b) associative retrieval is not an effective recall device but is rather a precision device in many cases, operating by increasing the weight of significant terms rather than by introducing new significant terms;

c) as a method of improving both precision and recall, a properly made thesaurus is generally preferable to associative procedures.

The usefulness of an associative method in operational retrieval experiments will probably be restricted to aiding dictionary constructors by pointing out unusual word relationships in a document collection, and as an alternative option for requests which do not perform well in a thesaurus. The method does not appear to be a reliable substitute for a thesaurus for the typical retrieval operation.

References

[1] V. Giuliano and P. Jones, Linear Associative Information Retrieval, in Vistas in Information Handling, P. Howerton, editor, Spartan Books, Washington, D. C. 1963.

[2] L. B. Doyle, Indexing and Abstracting by Association, American Documentation, Vol. 13, No. 4, 1962.

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

[4] G. Salton and M. Lesk, Computer Evaluation of Indexing and Text Processing, Journal of the ACM, Vol. 15, No. 1, January 1960.

[5] K. Sparck Jones and D. M. Jackson, The Use of the Theory of Clumps for Information Retrieval, Cambridge Language Research Unit, Cambridge, England 1967.