1986-ACM CONFERENCE ON RESEARCH
AND DEVELOPMENT IN

INFORMATION RETRIEVAL

Edited by FAUSTO RABITTI

September 8–10, 1986

1986 - ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN

INFORMATION RETRIEVAL

EDITED BY FAUSTO RABITTI

PALAZZO dei CONGRESSI
Via Matteotti, 1
PISA - ITALY

SEPTEMBER 8 - 10, 1986

Additional copies of these proceedings may be ordered prepaid from

ACM Order Department
P.O. Box 64145
Baltimore, Maryland 21264
ACM Order 606860
ISBN 0-89791-187-3

o from
ACM - CNR CONFERENCE
IEI - Via S. Maria, 46
56100 PISA, Italy

Price: member : $24
non-member: $18

Finito di stampare nel Luglio 1986 presso le Officine Grafiche
della Lito-Tipografia VIGO CURSI - Via S.Maria, 77 - Pisa.

PREFACE

The ACM Conference on Research and Development in Information Retrieval is being held this year in Pisa (Italy). This conference is traditionally held on alternate years in North America and in Europe. In 1985, it took place in Montreal (Canada) and next year it will be in New Orleans (USA). In Europe, it has already been held three times in England and once in West Germany; this is the first time in Italy.

This conference is intended to identify and encourage research, development and applications in Information Retrieval. Its principal objective is to provide an international forum to promote an understanding of current research, and to stimulate the exchange of ideas and experiences in information retrieval systems. The proceedings of the conference constitutes, year after year, an important record of the scientific and technical evolution in the area of information retrieval throughout the world.

The 1986-ACM Conference on Research and Development in Information Retrieval has been sponsored by the Italian National Research Council (CNR), a public organization which performs, coordinates and supports research activities in various fields. Luigi Rossi Bernardi, the President of CNR, is the chairman of the conference. Most of the organization has been carried out in Pisa at the "Istituto di Elaborazione dellInformazione" (IEI), an institute of CNR working in information processing. We should like to thank the Director of IEI, Franco Denoth, for his support and we are deeply grateful to all the staff of IEI who have contributed towards the organization of the conference. In particular, the efforts of Ettore Ricciardi, our treasurer, and of Manuela Mennucci, Anna Passerotti and Carol Peters have been vital to the conference.

We have benefitted from the collaboration of several important groups and organizations, such as the ACM Special Interest Group in Information Retrieval, the BCS-IRSG (British Computing Society), the ESA-IRS (European Space Agency), and, within Italy, the AICA-GLIR (the working group in I.R. of the Italian Computing Society, with the active cooperation of its chairman, Maristella Agosti) and IDI. The conference has also received a generous financial support from several companies, including IBM Italy, Siemens Data, Olivetti, Cerved, Sperry, Jackson and Logos.

We should like to express our appreciation to the members of the Conference Program Committee, and in particular to Gerard Salton, chairman the program committee for North America, who have all contributed to the selection of the papers contained in this book and which are being presented at the conference.

Fausto Rabitti

Conference Chairman

LUIGI ROSSI BERNARDI

President of CNR, Italy

Program Chairmen

FAUSTO RABITTI
IEI-CNR, Italy

GERARD SALTON
Cornell University, USA

Treasurer and Local Arrangements

ETTORE RICCIARDI, IEI-CNR, Italy

Program Committee

M. AGOSTI - Italy
P. BOLLMAN - W. Germany
Y. CHIARAMELLA - France
E. CHOURAQUI - France
F. NALDI - Italy
J.H. SCHEK - W. Germany
J. TAIT - U.K.
C.J. VAN RIJSBERGEN - U.K.
P. WILLETT - U.K.
R. ALLEN - USA
B. CROFT - USA
T. DOSZKOCS - USA
M. LESK - USA
C.T. YU - USA

Organizing Committee

F. RABITTI - IEI-CNR, Pisa
E. RICCIARDI - IEI-CNR, Pisa
M. AGOSTI - University of Padua
F. NALDI - SIAM-CNR, Milan
E. ONORATO - ESA-IRS
V. MUSSO - IDI
M. MORELLI - IBM Italia
C. MORTARINO - SIEMENS DATA
V. TOGNANA - CERVED
D. MONACO - OLIVETTI
G. CALDARA - SPERRY
F. CANAVESE - G.E. JACKSON
C. COCOLA - LOGOS

CONFERENCE SECRETARIAT

M. MANNUCCI, A. PASSEROTTI
IEI-CNR Via S. Maria, 46
56100 Pisa (Italy)
Tel. +39-50-500159 * Tlx 590305 IEICNR I

THIS CONFERENCE IS SPONSORED BY

CONSIGLIO NAZIONALE DELLE RICERCHE

IN COOPERATION WITH

ACM - SIGIR

A.I.C.A. - GLIR
Associazione Italiana per l'Informatica
ed il Calcolo Automatico

BCS - IRSG

idi INFORMAZIONE DOCUMENTAZIONE INDUSTRIA

ESA - IRS

THIS CONFERENCE IS SUPPORTED BY

Istituto di
Elaborazione
dell'Informazione

IBM ITALIA

Siemens Data

olivetti
Ing. C. Olivetti & C., S.p.A.

CERVED SOCIETÀ NAZIONALE DI INFORMATICA DELLE CAMERE DI COMMERCIO ITALIANE

SPERRY

GRUPPO EDITORIALE
JACKSON
Milano-San Francisco-Londra-Madrid

INDICE AUTORI

CONFERENCE PROGRAM

monday september 8, 1986

tuesday september 9, 1986

wednesday september 10, 1986

Recent Trends in Automatic Information Retrieval

Gerard Salton*
Department of Computer Science
Cornell University
Ithaca, NY 14853

Abstract

Substantial successes were achieved in the early years in automatic indexing and retrieval using single term indexing theories with term weight assignments based on frequency considerations. The development of more refined indexing systems using thesaurus aids and automatically constructed term association maps changed the retrieval effectiveness only slightly. The recent introduction of the relevance concept in the form of probabilistic retrieval models provided a firm basis for term weighting and document ranking practices. However, the probabilistic methods were not helpful in substantially enhancing the retrieval effectiveness.

At the present time, attempts are made to add artificial intelligence concepts to the document retrieval environment in the form of fancy graphics interfaces, learning systems for query and document indexing and for collection searching, extended logic models relating documents and information requests, and analysis methods based on the use of semantic maps and other kinds of knowledge structures. Using the earlier developments and evaluation results as guidelines, an attempt is made to outline the information retrieval environment of the future and to assess the usefulness of some of the currently proposed search and retrieval methods.

1. Automatic Information Retrieval

Information retrieval deals with text analysis, text storage, and the retrieval of stored records that are similar in some sense to information requests received from a population of users. From the beginning, it was realized that some retrieval tasks would be more difficult to mechanize than certain others: whereas the computation of similarity coefficients between the content identifiers attached to stored records and to user information requests could be mechanized relatively easily, it was argued early on that the content analysis task itself would probably have to be performed by human beings and not by machines for a long time to come:

> "it is very likely that manual indexing (content analysis) by cheap clerical labor will still, on average, be qualitatively superior to any kind of automatic indexing....neither the assignment of topic terms to a given request, nor the reformulation of a request are processes which could conceivably be adequately mechanized, contrary to some speculation in this direction." [BARH62]

In spite of these predictions, efforts have continued over the years to devise workable automatic indexing methods, and effective as well as efficient automatic retrieval procedures. In the early years, the goodness of a text word for content identification purposes was thought to be dependent principally on the frequency of occurrence of a word in a particular text:

> "a notion occurring at least twice in the same paragraph would be considered a major notion. A notion which occurs in the immediately preceding or succeeding paragraph would be considered a major notion even though it appears only once in the paragraph under consideration. Notations for major notions as just defined would then be listed in some standard order as representative of that paragraph..." [LUHN57]

More recently, it was realized that words occurring frequently in the texts of particular documents could not be used to distinguish these documents from the remaining texts of a collection if their occurrence frequency was high also in all the other available documents. This led to the notion that a good term should have a high term frequency in a particular document, but a low overall frequency in the collection. [SPAR72] These insights were used to generate a term weighting formula, known as "tf × idf", consisting of the product of the term frequency inside a given document multiplied by the inverse document frequency representing a function inverse to the number of documents to which a term is assigned. [SALT73]

Somewhat surprisingly, the single term indexing theories in which document or query content is represented by weighted sets of single terms extracted from the corresponding document or query

*This study was supported in part by the National Science Foundation under grant IST-83-16166.

texts proved to be quite powerful. Various experiments performed to compare manual content analysis with automatic single term indexing systems indicated in each case that quite simple automatic term extraction systems were not inferior in retrieval effectiveness to more conventional manual analysis methods. [SALT85a] A typical sample evaluation output is shown in Table 1 in terms of recall (the proportion of relevant materials retrieved) and precision (the proportion of retrieved materials that are relevant) for a collection of 44,000 document titles and abstracts in aeronautics used with 40 search requests. [CLEV77] The output of Table 1 shows that the manual (or intellectual) indexing system produced somewhat better search precision at the cost of substantially reduced search recall in comparison with the automatic term extraction system.

Content Analysis Method	Recall	Precision
Automatic natural language indexing (automatic text search of document titles and abstracts)	0.78	0.63
Controlled language manual indexing (indexing by trained personnel using a schedule of controlled terms	0.56	0.74

Comparison of Manual and Automatic Indexing
(recall and precision figures for
44,000 NASA documents in aeronautics
averaged for 40 search requests)

Table 1

Cleverdon, who was in charge of the NASA system test reaches the following conclusions based on the result of Table 1:

> "It appears impossible to reach any other conclusion than that, within the parameters of this test, natural language searching on titles and abstracts proved at least equal to, and probably superior to (manual) searching on controlled language terms." [CLEV77]

While the original single term automatic text analysis systems proved quite effective, it was obvious that the single term sets extracted from document texts could offer only a simplified picture of actual text meaning and content. The suggestion then arose that more refined content identifying units than single terms be used for content representation. In particular, it was believed that two types of vocabulary relationships should be taken into account in analyzing text content: [LYON68]

- the paradigmatic relations that cover term associations such as synonyms and hierarchical term inclusion relations, that always exist between particular terms, regardless of the context in which the terms are used (for example, the relation between "computer" and "calculator", or the relation between "Paris" and "France");

- the syntagmatic relations between terms in which the relationship between linguistic entities depends on the context in which the terms are used (for example, the relation between "united" and "states" in the phrase "United States".

Various methods are outlined in the literature designed to make use of paradigmatic and syntagmatic relations in text content analysis. The main possibilities consist in using a thesaurus of related terms, or alternatively in constructing so-called term association maps that are capable of identifying "associated" term sets, or finally in generating complex text identifying units such as term phrases consisting of combinations of single terms. The idea in each case is to take particular single terms and to use them as entry points for the identification of related notions from the thesaurus or from the term association map; alternatively the single terms are used as a starting points for the generation of phrases. The new, associated terms and phrases can then be added to the originally available content identifiers, or the new terms can replace originally available terms.

When the subject area is narrowly circumscribed, and knowledgeable subject experts are available, useful thesaurus arrangements can be manually constructed by human experts. Such thesauruses may then provide substantial enhancements in retrieval effectiveness. Table 2 shows the average search precision obtained at certain fixed recall points for a collection of 400 documents in engineering used with 17 search requests. In Table 2, the performance obtained by extracting thesaurus class entries from a manually constructed thesaurus (the Harris Three thesaurus) is compared with a simpler analysis system where the weighted terms obtained from query and document texts are used directly for content analysis purposes. The single term base case of Table 2 uses a term frequency (tf) term weighting system which normally performs better than an unweighted single term system, but not as well as the preferred tf × idf term weights.

The output of Table 2 shows that at the high recall end of the performance range, the thesaurus provides much better retrieval output than the weighted word stem process. The average advantage of the thesaurus when low recall points are also taken into account is 13 percent. [SALT68a] The use of thesauruses is widely advocated as a means for normalizing the vocabulary of document texts. However, the construction of useful thesauruses is an art rather than a science and requires extensive knowledge of the particular subject area under consideration and of the record collections to be processed. The construction of a thesaurus is thus a major undertaking which needs to be repeated for each distinct subject area. The so-called thesaurus method is therefore difficult to implement efficiently in operational environments.

Recall	Average Search Precision				
	Weighted Word Stems	Harris Three Thesaurus		Automatic Term Associations	
.1	.9563	.9735	+ 2%	.7385	-23%
.3	.7986	.8245	+ 3%	.5844	-27%
.5	.6371	.7146	+11%	.5187	-19%
.7	.4877	.6012	+19%	.4452	- 9%
.9	.3426	.4973	+31%	.3794	+11%
Average Improvement			+13%		-13%

Sample Thesaurus and Associative
Indexing Performance
(IRE collection, 400 documents, 19 queries)

Table 2

In the early years it was hoped that thesauruses could be built automatically by studying the occurrence characteristics of the terms in the documents of a collection, and grouping into common thesaurus classes those terms that would co-occur sufficiently often in the texts of the documents. Later it was recognized that thesauruses constructed by using the occurrence characteristics of the vocabulary in the documents of a collection do not in fact identify generally valid paradigmatic term relations, but provide instead locally valid syntagmatic relations derived from the vocabulary of the collections under consideration. [LESK69] Such locally valid term grouping systems are known as term association maps, and the indexing method which consists in adding associated terms obtained from a term association map to the originally available document terms is known as associative indexing. [DOYL61,GIUL62]

In practice, it is found that the use of term associations can improve the search recall by providing new matches between the terms assigned to queries and documents that were not available before the identification of the associated terms. In addition, the search precision can also be enhanced by reinforcing the strength of already existing term matches. [LESK69] Unfortunately, the experimental evidence indicates that only about 20 percent of automatically derived associations between pairs of terms are semantically valid. As a result, the associative indexing process provides many erroneous associations in addition to some useful ones and on average the associative methodology does not produce guaranteed advantages in retrieval effectiveness.

The right hand side of Table 2 contains evaluation results for the associative indexing method applied to the document collection in engineering previously used with the Harris Three thesaurus. The results of Table 2 show that the automatic term associations provide an increase in average search precision only at the high recall end of the performance range. Overall, the average search precision of the associative method decreases by 13 percent compared with the use of weighted single terms using term frequency weights. [SALT68b, p.130]

The use of manually constructed thesauruses and automatically derived term association maps is designed principally to enhance search recall by adding new related terms to the terms originally available to identify document and query content. Search precision may be enhanced by using narrow, specific content identifiers instead of broader, less specific terms. This suggests that broad single terms, exhibiting high occurrence frequencies in the documents of a collection be replaced by narrower term phrases. For example, a broad term such as "computer" might be replaced by narrower constructions such as "computer system", or "computer programming".

In principle, the identification of useful term phrases must be based on syntactic language analysis systems, supplemented by semantic know-how for the subject area under consideration. Unfortunately, complete linguistic analysis methods covering topic areas of reasonable scope are not available for operational use. In practice, it is therefore necessary to fall back to simpler methodologies in which phrases are identified as sequences of terms that co-occur in certain contexts in the documents and queries of a collection. Because phrase formation is a term narrowing operation, one normally insists that one component of the phrase (the phrase head) be a high-frequency component in the collection under consideration. Furthermore, a domain of co-occurrence must be specified between the phrase components that are to be included in a phrase: typically, a phrase may be formed when two particular terms occur in the same sentence of a document or query, and when the number of intervening words does not exceed a stated small number.

The evaluation results for a typical nonsyntactic phrase formation process are shown in Table 3 for two collections of documents in computer science (the CACM collection) and documentation (the CISI collection) averaged over 52 and 76 user queries respectively. The base run for the output of Table 3 is a single term assignment process where the terms are assigned tf × idf weights. The phrase process used for the experiments of Table 3 consisted in adding phrases to the available single terms whenever pairs of adjacent word stems would occur in the same sentences of documents and queries. The phrase weight used is, in each case, the average of the normalized tf × idf weights of the individual phrase components. [FAGA85]

The output of Table 3 shows that the phrase formation process operates as expected for the CACM collection since the narrowing of the indexing vocabulary enhances the search precision at the low recall end of the performance range. Overall the advantage of the phrase formation process is about 8 percent in average precision at the 5 recall points for the CACM collection. The same phrase process unfortunately does not work well for the CISI collection in documentation; in that case the performance with and without phrases is approximately the same. For CISI, the presence of good, semantically acceptable phrases that might serve in improving retrieval performance is evidently compensated by the formation of many adjacent matching word groups that are extraneous and hence are unable to pull out relevant materials when they happen to be present both in the queries and documents of a collection.

Recall	CACM (3204 docs, 52 queries)			CISI (1460 docs, 76 queries		
	Single Terms	Single Terms plus Phrases		Single Terms	Single Terms plus Phrases	
.1	.5086	.5580	+10%	.4919	.4813	- 2%
.3	.3672	.4065	+11%	.3118	.3158	+ 1%
.5	.2398	.2835	+18%	.2320	.2291	- 1%
.7	.1462	.1466	0%	.1504	.1463	- 3%
.9	.0711	.0704	- 1%	.0739	.0717	- 3%
Average Improvement			+7.6%			-1.6%

Average Search Precision for Single Term Indexing with Nonsyntactic Phrase Indexing for Four Sample Document Collections

Table 3

The evaluation results presented in Tables 1 to 3 indicate that the available automatic term extraction and term weighting methods are surprisingly effective in retrieving a large proportion of the relevant materials included in various document collections, and in rejecting a large proportion of the nonrelevant items. In general, the single term automatic indexing methods are easily competitive with manual content analysis methods based on the intellectual efforts of trained indexers. On the other hand, attempts to improve the recall performance of the single term methods by adding related terms derived from thesauruses and word association maps, or to enhance the precision by adding automatically determined term phrases, are much less successful.

The procedures designed to generate the complex indexing units in the form of thesaurus classes or term phrases are obviously not sufficiently discriminating to insure that a large proportion of these more refined content identification units are in fact appropriate for the collections under consideration. The manual thesaurus construction process is not easily replicated for different subject areas, and the phrase formation process may require additional syntactic and semantic controls to insure that the number of extraneous identifiers that are detected remains limited. For present purposes it appears reasonable to stay with the simpler single-term indexing methods.

2. The Introduction of Term Relevance

The notion of term relevance is absent from the methods examined in the previous section. That is, no distinction is made between terms that occur in documents that are either relevant, or nonrelevant, to particular queries. Instead the term values are determined by the term occurrence characteristics in the whole document collection regardless of the relevance properties of the individual items.

A number of retrieval models have been introduced in which the concept of document or term relevance takes a central place. The best known of these is the probabilistic retrieval model. [BOOK75, COOP78, HART75, KRAF78, ROBE83, VANR79] In the probabilistic model, the retrieval operation consists in ranking the documents of a collection in decreasing order of the ratio of their probability of being relevant to a query to the probability of their being nonrelevant. Given a document $\underline{x}$, it then becomes necessary to estimate the probabilities $P(\underline{x}|rel)$ and $P(\underline{x}|nonrel)$. These probabilities can in turn be made to depend on the probabilities of relevance of the individual terms x_i occurring in the respective documents. When one assumes that the individual terms occur independently of each other in the relevant and the nonrelevant documents of a collection, an optimum term weight w_i can be derived for each term x_i occurring in an information request which is equal to

$$w_i = \log \frac{P(x_i|rel)[1-P(x_i|nonrel)]}{P(x_i|nonrel)[1-P(x_i|rel)]} \qquad (1)$$

Assuming that the individual term relevance weights, w_i are available, each document can then be assigned a global document relevance score equal to the sum of the w_i factors for all query terms that are also present in each particular document. This provides the needed document ranking function.

The probabilistic retrieval model is attractive because it provides a theoretical foundation for the retrieval operation which takes into account the notion of document relevance. Furthermore, it is possible in the probabilistic environment to take into account at least some of the dependencies and relationships between the terms used to identify the queries and the stored records. [HARP78, VANR77] Finally, the probabilistic model offers justification for various methods that had previously been used in automatic retrieval environments on an empirical basis. For example, the useful inverse document frequency (idf) term weighting system, which ranks the terms in decreasing order of the total number of documents in a collection to which the terms are assigned, can be shown under appropriate assumptions to be similar to the term relevance weight of expression (1). [WUSA81]

Unfortunately, the probabilistic methodology has been disappointing in practice for a variety of reasons.

- the basic probabilistic methodology rests on the use of initially unweighted document and query terms, even though it is known that substantial differences exist in the worth of individual terms for content identification purposes;

- the computation of the term relevance weights depends on knowledge about the relevance of individual documents to information requests; since this relevance information is normally not available at search time, the probabilistic weights can be estimated only imperfectly;

- too many subsets of terms with potential dependencies between them exist to make it practical to include all of them in any retrieval model; it is then necessary to use simplified term dependency models from which some useful information will necessarily be missing.

The result is that the probabilistic model has been important mainly for providing solid foundations for much of the information retrieval work rather than as a practical tool in experimental or operational situations.

This last point may be illustrated by considering the results of term weighting experiments carried out recently by Christopher Buckley at Cornell University, where 154 different term weighting methods were processed against six different document collections. [BUCK85] The different weighting system were obtained by considering various forms of the term frequency factor (tf) of a term, the inverse document frequency factor (idf) of the term, and the normalization method, if any, used to compute the total term weights. In practice, different term weighting systems are usable for query terms and document terms, and the normalization factor applied to the weights may be used to simulate various query-document matching methods.

Consider a particular document $D = (w_{d_1}, w_{d_2}, \ldots, w_{d_t})$ and query $Q = (w_{q_1}, w_{q_2}, \ldots, w_{q_t})$, where w_{d_k} and w_{q_k} represent the weights of the kth terms in D and Q respectively. The similarity between D and Q may then be computed as

$$sim(D,Q) = \sum_{k=1}^{t} w_{d_k} \cdot w_{q_k} \quad (2)$$

For unnormalized term weights, expression (2) represents simply the vector product between the vector representations of D and Q. On the other hand, when a cosine normalization is used for w_{d_k} and w_{q_k}, then expression (2) becomes a cosine computation. For example, when standard term frequency weights are in use, $\sum_{k=1}^{t} tf_{d_k} \cdot tf_{q_k}$ represents a normal weighted vector product. However, with weighted term frequency factors $tf_{d_k} / \sqrt{\sum_{i=1}^{t} tf_{d_{k_i}}}$ and $tf_{q_k} / \sqrt{\sum_{i=1.}^{t} tf_{q_i}}$, the formula of expression (2) produces a cosine measure.

Table 4 shows the average precision results for 6 particular term weighting systems, including two probabilistic weighting systems (termed best probabilistic term weighting and probabilistic binary term independence, respectively), three standard term weighting methods (termed begt tf × idf, normal idf, and normal tf, respectively), and finally a coordination level match corresponding to the absence of any weighting before using expression (2) to compare queries and documents. For the probabilistic term weights, approximations are used which make it possible to compute the formula of expression (1) without knowing any relevance information [CROF79]. The run corresponding to "best probabilistic weight" uses the normal probabilistic approximation for query terms (that is, $\log(N-n_i)/n_i$, where N is collection size and n_i is the number of documents with term i), and a term frequency weight of limited range defined as (0.5 + 0.5 tfi/max tf) for document terms. The query-document similarity measure for the "best probabilistic weight" system is then computed as

$$sim(D,Q) = \sum_{k=1}^{t} \left[0.5 + \frac{0.5\ tf_{d_k}}{\max tf_d}\right] \log\left(\frac{N-n_{q_k}}{n_{q_k}}\right) \quad (3)$$

Analogously the weighting factors being multiplied for the "best (tf × idf)" run are respectively

$$\frac{\left(tf_{d_k} \cdot \log\frac{N}{n_{d_k}}\right)}{\sqrt{\sum_{k=1}^{t}\left(tf_{d_k} \log\frac{N}{n_{d_k}}\right)^2}} \quad \text{and} \quad \left(0.5 + \frac{0.5\ tf_{q_k}}{\max tf_q}\right) \log\frac{N}{n_{q_k}}. \quad (4)$$

In addition to the average precision values for the six weighting systems, Table 4 also shows the ranks out of 154 for the given weighting methods used with each collection. In each case, a rank of 1 designates the best weighting system for each collection, and 154 the worst. The result of Table 4 indicates that the (tf × idf) weighting systems are generally better than the best probabilistic methods; these in turn are better than the standard idf weighting systems, the conventional binary term independence methods, the normal tf methods, and finally the coordination level matches, in that order. The average rank of the best (tf × idf) method is 17.8 for the six collections whereas the average rank of the coordination level match is 117.1. An alternative weighted probabilistic system proposed by Croft does not produce result that are as good as those of the best probabilistic system of expression (3). [CROF81,CROF83]

The results of Table 4 show that the basic ranking of the various weighting systems is the same for all collections with the exception of the NPL collection where the probabilistic weighting systems are better than the other frequency based weights. Experimental results based on the NPL collection were used earlier to claim superiority for the probabilistic methodology. [CROF84] However, as Table 4 shows, the NPL records represent a special case. The NPL document and query representations are very short and a large mass of low frequency terms are used to index the collection. In fact, most NPL terms exhibit a frequency of 1 in the collection; the normal term frequency weights are

Collection	Best tf × idf		Best probabilistic term weighting		Normal idf	
	Rank	Average Precision	Rank	Average Precision	Rank	Average Precision
CACM 3204	7	.3130	34	.2713	55	.2385
CISI 1460	11	.1874	73	.1377	86	.1314
MED 1033	5	.5628	26	.5443	84	.5038
CRAN 1398	22	.3821	27	.3772	98	.3174
INSPEC 12684	7	.2626	42	.2092	72	.1779
NPL 11429	55	.1933	2	.2755	25	.2401
Average rank	17.8		34.0		70.0	

Collection	Probablistic binary term independence		Normal tf		Coordination Level	
	Rank	Average	Rank Precision	Average	Rank Precision	Average
CACM 3204	78	.2197	93	.1910	104	.1721
CISI 1460	99	.1237	109	.1196	134	.1019
MED 1033	75	.5072	119	.4641	148	.4070
CRAN 1398	100	.3154	74	.3371	147	.2370
INSPEC 12684	92	.1549	83	.1621	130	.0970
NPL 11429	12	.2583	83	.1749	40	.2094
Average rank	76.0		93.5		117.1	

Average Precision for Various Term Weighting Systems
for Six Document Collections

Table 4

therefore not effective in that case. The best probabilistic term weighting system is ranked 34th out of the 154 methods considered.

Another possibility for introducing the effect of document and term relevance into the retrieval process is the well-known relevance feedback process. [ROCC71,IDEE71] In that case, the results of an initial search run are used automatically to reformulate the search request by increasing the weights of query terms that are present in previously retrieved documents termed relevant to the query, and contrariwise in decreasing the weights of query terms also present in the nonrelevant documents previously retrieved. In addition to changing the query term weights, the queries can also be expanded by adding new terms to the query formulations taken from the relevant documents previously retrieved. The relevance feedback process is normally applied to vector queries where each query is formulated on a set of query terms. However methods are also available for using relevance feedback with Boolean queries. [SALT85b]

Table 5 shows typical relevance feedback results for one iteration of feedback -- that is, initial search followed by a single query reformulation -- for the two collections previously used for the experiments of Table 3. As usual, average precision results are shown for the 52 CACM queries and the 76 CISI queries for feedback systems with and without query term expansion. The base run used for comparison purposes is a (tf × idf) term weighting system with a cosine similarity normalization used for query and document terms. Fifteen initially retrieved documents are examined for relevance in the runs of Table 5, and all relevant retrieved items plus the top nonrelevant retrieved item are used for feedback purposes (the so-called "Dec Hi" method [IDEE71]).

The output of Table 5 shows that one iteration of relevance feedback produces over 90 percent improvement in average precision for CACM and nearly 50 percent for CISI, thus confirming the usefulness of query reformulation methods that are based on simple user input such as relevance judgments for previously retrieved items. When relevance information is available, useful performance improvements are thus obtainable with quite simple iterative search systems.