Combining Automatic and Manual Index Representations in Probabilistic Retrieval *

T.B. Rajashekar
National Centre for Science Information
Indian Institute of Science
Bangalore 560 012, India

W. Bruce Croft
Computer Science Department
University of Massachusetts
Amherst, MA 01003

Abstract

Results from research in information retrieval suggest that significant improvements in retrieval effectiveness could be obtained by combining results from multiple index representations and query strategies. Recently, an inference network based probabilistic retrieval model has been proposed, which views information retrieval as an evidential reasoning process in which multiple sources of evidence about document and query content are combined to estimate the relevance probabilities. In this paper we report the results of a series of experiments we conducted using a specific implementation of this model to study the retrieval effectiveness of combining automatic and manual index representations in queries and documents. The results indicate that significant benefits in retrieval effectiveness can be obtained through combined representations.

Keywords: Information storage and retrieval, inference networks, probabilistic retrieval.

*This work was supported in part by a UNDP fellowship held by the first author at the University of Massachusetts, Amherst, U.S.A. during Sept 1992 to Feb 1993. Additional support was provided by the NSF Center for Intelligent Information Retrieval at Amherst.

1 Introduction

With the enormous growth in the number and size of bibliographic, full text and other electronic information sources, information service providers are under constant pressure to provide their users with the most relevant items of information, in response to their information needs. Unfortunately, many of the currently operational information retrieval systems do not respond to this requirement adequately. Each information item in these databases has several clues (properties or content representations) about relevance in the form of natural language text (e.g., title, abstract, full text), manually assigned index terms, subject categories, etc. Similarly, a variety of clues can be obtained from the users about their information needs (e.g., natural language descriptions, term importance, known relevant papers, etc.). This information is typically used for the construction of Boolean queries. In Appendix 1, the relevant portions of a completed user profile information sheet used in a SDI service are shown, to illustrate the way this information is obtained (Rajashekar, 1988). Conventional retrieval systems do not make full use of this information due to the exact match interpretation of Boolean queries. While these systems are quite effective for some kinds of searching (e.g., known-item searching), when it comes to more general searching or for untrained users, they often result in either no output, not enough output, or too much output (Cooper, 1988; Maron, 1988).

The reason for the poor average performance of these systems lies in the binary (or strict) interpretation of the relationship between document properties and relevance. This relationship is more appropriately interpreted as probabilistic (Maron, 1988). Several "best-match" retrieval models have been proposed which compute a Retrieval Status Value (RSV) for each document based on the degree or probability of relevance of the document to the query, and rank the retrieved documents by their RSVs. The best known of such models are the vector space and probabilistic retrieval models (Salton & McGill, 1983; Bookstein, 1985; Belkin & Croft, 1987; Turtle & Croft, 1990). These models support natural language query processing and often employ automatic, non binary indexing of documents and/ or queries, for computing the RSVs. Best-match retrieval techniques have almost always performed much better than the exact-match techniques under laboratory conditions using test collections of a few thousand records, and we are beginning to witness the commercialization of these techniques and evaluation of their effectiveness with large text databases (Wagers, 1992; DARPA, 1990; Harman & Candela, 1991; Callan et al., 1992).

The results of some recent information retrieval research indicate the possibility of further improvements in retrieval effectiveness. These results show that a) a given query will retrieve different documents when applied to different representations, even when the average retrieval performance (recall/ precision) achieved with each representation is the same, indicating that documents retrieved by multiple representations are more likely to be relevant (Katzer et al., 1982; Croft & Harper, 1979; Fox et al., 1988), and b) given a single natural language description of an information need, different searchers will formulate different queries to represent different aspects of that need and will retrieve different documents, even when average performance is the same for each searcher, indicating that documents retrieved by multiple searchers (search strategies) are more likely to be relevant (Katzer et al., 1982; McGill et al., 1979; Croft, 1987).

These results indicate that significant improvements in retrieval effectiveness may be possible if we can combine results obtained by using multiple document representations and query strategies. By adapting retrieval techniques that support this capability, operational retrieval systems can better exploit the variety of document and query clues that already exist.

Recently, an inference network-based probabilistic retrieval model has been proposed which views information retrieval as an evidential reasoning process in which multiple sources of evidence about document and query content are combined to estimate the probability that a given document matches an information need (Turtle & Croft, 1990). Different representations of the document content, different representations of the information need, and domain knowledge such as thesaurus can all be taken into account under this model. A retrieval system INQUERY has recently been developed based on this model, supporting sophisticated indexing and complex query formulation (Callan et al., 1992). INQUERY has been used successfully on a variety of document and text databases ranging up to a few gigabytes in size.

In the study reported in this paper, our research goal was to demonstrate the flexibility that a probabilistic retrieval model provides an information retrieval system designer or service provider to combine manual and automatic index representations in documents and user queries, to offer significantly improved services to users.

In more specific terms, we demonstrate this through a series of experiments we conducted using INQUERY to evaluate the following hypotheses :

1. An information need, expressed in a single index representation, will improve re-

trieval effectiveness when multiple sources of evidence are available in documents for inference.

2. Significant improvements in retrieval effectiveness can be obtained by combining results from multiple index representations in queries.

3. Significant improvements in retrieval effectiveness can be obtained by combining results from multiple query strategies.

Our interest here is in index representations for subject access like controlled vocabulary terms, classification codes, subject headings, indexer selected terms and phrases from natural language text (keywords), automatically generated index terms, etc. By query strategy we mean the query syntax and its interpretation. For example, the syntax of a query can be Boolean, but its interpretation could be exact-match or probabilistic. Other examples of query strategies are natural language query, weighted term query, etc.

In section 2 we briefly describe the probabilistic inference net model which is the basis of our experiments. In section 3 we describe the experimental methodology. In section 4 we present the experimental results and a discussion of these results. Finally, in section 5, we draw conclusions from these results.

2 Probabilistic Inference Network Retrieval Model

The experiments described in sections 3 and 4 were carried out using INQUERY, a new generation retrieval engine developed at the Information Retrieval Laboratory in the University of Massachusetts. It is a specific implementation of the probabilistic inference net retrieval model. In what follows, we give a brief description of this model and the INQUERY operators used for this study. More details of INQUERY can be found in (Callan et al., 1992) and the inference net model in (Turtle, 1990; Turtle & Croft, 1991a; Turtle & Croft, 1991b). In this paper, the emphasis will be on the ability of the model to handle multiple sources of evidence.

The inference net model is a probabilistic retrieval model in that it follows the Probability Ranking Principle. A probabilistic model calculates P(Relevant|Document,Query), which is the probability that a user decides a document is relevant given a particular document and query (Robertson, 1977). The inference net model takes a slightly different approach in that it computes P(I|Document), which is the probability that a user's information need

is satisfied given a particular document. The inference net model is based on Bayesian inference networks (Pearl, 1988). These are directed, acyclic dependency graphs (DAG) in which nodes represent propositional variables or constants and edges represent dependence relations between propositions. If a proposition represented by a node p "causes" or implies the proposition represented by node q, we draw a directed edge from p to q. The node q contains a matrix (a *link* matrix) that specifies $P(q|p)$ for all possible values of the two variables. When a node has multiple parents, the matrix specifies the dependence of that node on the set of parents and characterizes the dependence relationship between that node and all nodes representing its potential causes. Given a set of prior probabilities for the roots of the network, these networks can be used to compute the probability or degree of belief associated with all remaining nodes.

Fig. 1 shows the basic document retrieval inference network used in INQUERY. It consists of two component networks : one for documents and one for queries. The document network is built once for a collection and its structure does not change during query processing. It consists of document nodes (d_j's) and concept representation nodes (r_m's). The content representation nodes or representation nodes can be divided into several subsets, each corresponding to a single representation technique that has been applied to the document texts. For example, if the phrase "information retrieval" has been extracted automatically and "information retrieval" has been manually assigned as an index term, then two representation nodes with distinct meanings will be created. We represent the assignment of a specific representation concept to a document by a directed arc to the representation node.

Each representation node contains a specification of the conditional probability associated with the node given its set of parent nodes. This specification incorporates the effect of any indexing weights (e.g., term frequency for each parent text) or term weights (e.g., inverse document frequency) associated with the representation concept. While, in principle, computation of this probability would require $O(2^n)$ space for a node with n parents, since only one document is instantiated to *true*, the link matrix at the representation nodes reduces to a simple canonical form. This probability estimate is very similar to the *tf.idf* weights used in many previous IR experiments (Salton & McGill, 1983).

The query network is an "inverted" DAG with a single leaf node (I) representing the user's information need, one or more query representations (q_k's) and multiple roots that correspond to the concepts that express the information need. The q_k nodes are intermediate

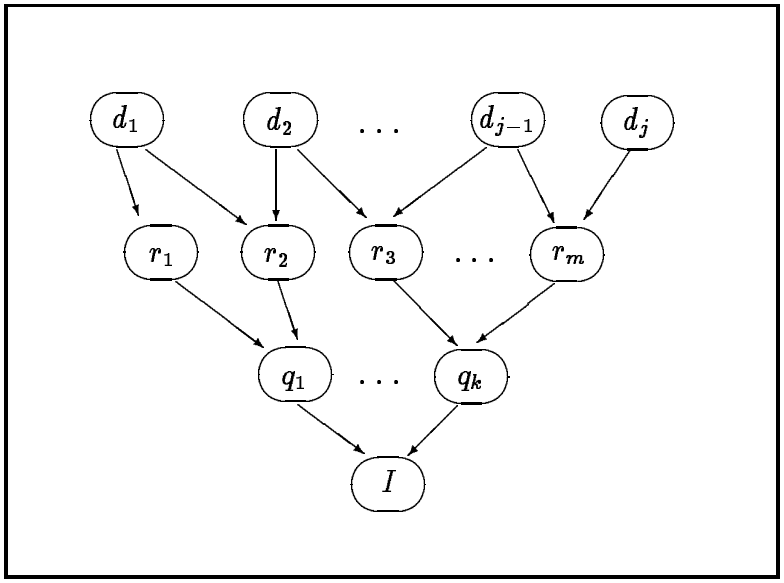

Figure 1: Basic document inference network

query nodes which may be used in cases where multiple query representations are used to express the information need (e.g., a Boolean combinations of representation nodes, a natural language description and a sample document).

A query is processed by constructing the query network and attaching it to the document network. The attachment of the query concept nodes to the document network has no effect on the structure of the document network. The probability that the information need is met given a particular document d_j is computed by instantiating d_j and computing the beliefs associated with each node in the query network, given this evidence. Beliefs are computed using the indexing weight specifications (term frequency, inverse document frequency, etc.) associated with the representation nodes. This is repeated for each document in the network and the probabilities used to rank the documents. Simplifying assumptions made for efficient implementation of inference networks, their construction and evaluation, are discussed in (Turtle & Croft, 1991a).

INQUERY uses several closed-form expressions to compute the belief associated with each node in the query network, given the beliefs in its parent nodes. This helps INQUERY in restricting the ways in which evidence is combined. The derivation of these expressions and their use are given in (Turtle, 1990). For a query node Q with parents $P_1,\ldots,P_n$ where $\mathrm{P}(P_1 = \mathit{true})=p_1,\ldots,\mathrm{P}(P_n = \mathit{true})=p_n$, the closed-form expressions are :

OPERATOR	ACTION
#and	AND the terms in the scope of the operator.
#or	OR the terms in the scope of the operator.
#not	NEGATE the term in the scope of the operator.
#sum	Value is the mean of the beliefs of the nodes.
#wsum	Value is the sum of weighted beliefs of the children scaled by the ratio of the weights of the parent node and sum of the children nodes.
#max	The belief is the maximum belief of all the parent nodes.
#n	Evaluate terms in the scope of the operator that occur within n words of each other in the text. The order of the terms does not matter for the first two terms.
#phrase	Evaluate the terms (or subsets of terms) occurring within three words of each other in the text. A "#3" operator is applied to the entire phrase and to each contiguous sub-phrase. The maximum over all these values is the final belief.
#syn	Synonym operator.

Table 1: The operators in the INQUERY query language.

$$\text{bel}_{\text{not}}(Q) = 1 - p_1 \quad (1)$$

$$\text{bel}_{\text{or}}(Q) = 1 - (1 - p_1) \cdot \ldots \cdot (1 - p_n) \quad (2)$$

$$\text{bel}_{\text{and}}(Q) = p_1 \cdot p_2 \cdot \ldots \cdot p_n \quad (3)$$

$$\text{bel}_{\text{max}}(Q) = \max(p_1, p_2, \ldots, p_n) \quad (4)$$

$$\text{bel}_{\text{wsum}}(Q) = \frac{(w_1 p_1 + w_2 p_2 + \ldots + w_n p_n) w_q}{(w_1 + w_2 + \ldots + w_n)} \quad (5)$$

$$\text{bel}_{\text{sum}}(Q) = \frac{(p_1 + p_2 + \ldots + p_n)}{n} \quad (6)$$

A user can pose queries to INQUERY by using either natural language or a structured query language. The INQUERY operators (see Table 1) permit the user to provide structural information in the query, including phrase and proximity requirements. Natural language queries are converted to the structured query language by applying the #sum operator to the terms in the query. It should be noted that the Boolean queries are evaluated in the probabilistic sense, and the documents ranked accordingly.

Henceforth in this paper, we will use 'index representation' to mean both 'content or representation type' and 'query concepts', a terminological difference used in this section for presentation convenience. The context of usage will clarify whether we are referring to

'index representations' in queries or documents. Furthermore, we will use 'query strategy' and 'query representation' synonymously.

The operators given in Table 1 provide a powerful means for combining different index representations and query strategies. For example, Boolean and natural language searches can be combined using the #sum operator as,

#sum((*Boolean Query*) #sum(*Natural language query*))

Similarly, different index representations of a query, for example using controlled vocabulary terms, keywords manually identified from the query and the full natural language query, can be combined as,

#sum(#sum(*Thesaurus terms*) #sum(*Keywords*) #sum(*Natural language query*))

Furthermore, the #wsum operator can be used to assign weights to individual query or index representations within these combined representations.

3 Experimental Methodology

The experiments reported in this paper were carried out using standard IR methodology in which a test collection consisting of text documents, natural language queries and relevance judgements for each query, is used to generate recall/ precision figures (Sparck Jones & van Rijsbergen, 1976). Comparisons of retrieval effectiveness are made using tables of precision values at ten standard recall points (i.e., 10,20,...,100), averaged over a set of queries, for each of the query/ index representations and their combination being evaluated. When two tests are being compared, we show the difference as the percentage change from the baseline test. A difference of 5 percent in average is generally considered significant, and a 10 percent difference is considered very significant (Sparck Jones & Bates, 1977).

A primary requirement for this study was the availability of a test collection supporting multiple representation types. We selected the INSPEC test collection (Salton & Buckley, 1988) for this study as it supports multiple document representations - controlled vocabulary terms, keywords (indexer selected significant terms and phrases from document titles and abstracts) and the natural language text of titles and abstracts themselves. The INSPEC subject categories would have been an interesting additional index representation to use for this study, but unfortunately the test collection records did not include this representation. The INSPEC test collection contains 12,684 records covering the areas of computer,

electrical and electronic engineering, 84 queries in natural language and standard relevance judgements. Out of the 84 queries in the test collection, we selected 50 queries for this study. This selection was made based on the clarity of their expression, enabling accurate identification of key concepts and construction of various query strategies required for the study. Selecting the INSPEC test collection for this study has the additional advantage that it is representative of the bibliographic databases used in many of the operational information retrieval service centres, and the observations and conclusions reached in this study would therefore be that much more appropriate for such settings.

We first generated the following basic automatic and manual index representations and query strategies required for this study :

A. Index Representations :

1. Queries :
   - Automatic index representation : Indexing each significant word in the query (*NLQ*). Note that the indexing is carried out at search time.
   - Manual index representation : 1) Analysis and representation of query concepts using thesaurus terms (*Th*), and 2) Analysis and representation of query concepts using keywords, i.e., terms and phrases manually identified from the natural language query (*KW*). During this analysis, the keywords were also assigned weights which were used to formulate weighted term queries (see below).
2. Documents :
   - Automatic index representation : Indexing each significant word in the title and abstract text fields (*Tx*).
   - Manual index representation : Indexing each word in thesaurus terms (*Th*) and keywords (*KW*).

B. Query Strategies :

1. Natural language query formulated as a probabilistic query using the #sum operator (*NLQ*).
2. Boolean query, formulated using keywords and the Boolean operators #and, #or and #not (*BOOL*).

3. Weighted term query, formulated as a probabilistic weighted sum query using keywords and the #wsum operator (*WTERM*). Two sets of weighted term queries were generated using a scale of two and three importance levels - most important (1.0) and less important (0.5), and most important (1.0), moderately important (0.5) and less important (0.3).

It may be noted that within the index representations in queries, multi-word terms in keywords and thesaurus terms were represented as phrases using the #phrase operator. In the subsequent sections of the paper we will use the abbreviations shown inside the brackets as short hand notation to refer to their respective representations.

It may be seen that many of operational information retrieval systems either support some or all of these index and query representations or possess enough details from which these representations can be easily generated (see, for example, Appendix 1). These individual representations are combined to generate specific combinations required for evaluating the research hypotheses of this study. Details of specific combinations produced are discussed in Section 4. We conducted three sets of experiments corresponding to the three hypotheses. In the first set of experiments, we compared the performance of single index representations in queries (Th, KW, NLQ) on the document file indexed on one, two and three sources of evidence (Th, KW, Tx). In the second set of experiments, while keeping the sources of evidence in the document file the same (a combination of Th, Tx and KW), we compared the performance of combined index representations in queries generated by a combination of two (Th,NLQ; Th,KW; NLQ,KW) and three index representations (Th,NLQ,KW). In the third set of experiments we compared the performance of individual query strategies (NLQ, BOOL, WTERM) with their combined representation (NLQ,BOOL; NLQ,WTERM). The results of these experiments are presented in the following section.

4 Experimental Results

We discuss the results in terms of the three hypotheses.

Hypothesis 1 : An information need, expressed in a single index representation, will improve retrieval performance when multiple sources of evidence are available in documents for inference.

	Document Index Files (Collection Size : 12684 docs.)						
	Th	KW	Tx	Th,KW	Th,Tx	KW,Tx	Th,KW,Tx
Unique Stems	1851	9722	17983	9840	18068	18323	18383
Max stem frequency	3162 (comput)	4821 (system)	11833 (system)	7508 (system)	14520 (system)	16654 (system)	19341 (system)
Stem occurrences	61872	144146	579290	206018	641162	723436	785308
Postings	56975	125889	417592	155119	444834	426836	450342
Max within doc freq	8	10	32	10	32	32	32

Table 2: Summary of collection statistics

To test this hypothesis, we used three query files, each file consisting of the 50 queries represented in a specific index representation. Queries in the first two files were formulated using the manual index representations 'Th' and 'KW' and the third file consisted of the natural language queries (NLQ) providing the automatic index representation. Probabilistic sum (operator #sum) was used as the search strategy.

Seven document inference network files were generated using the INSPEC records :

1. Three files of single source of evidence - Thesaurus (Th), keywords - indexer selected terms and phrases from title and abstracts (KW) and natural language text (titles and abstracts) (Tx),

2. Three files of two sources of evidence - Th,KW; Th,Tx; KW,Tx, and

3. A combined file of all the three sources of evidence - Th,KW,Tx.

The collection statistics for these seven index files is shown in Table 2.

Each of the three query files was processed on the corresponding single evidence (e.g., 'Th' query file on 'Th' index file), two evidence (e.g., 'Th' query file on 'Th,Tx' and 'Th,KW' index files) and the combined three evidence (e.g., 'Th' query file on 'Th,Tx,KW' index file) index files, and the results compared with the standard relevance judgements. A summary of the results obtained is given in Table 3. The figures shown are average precision obtained over ten standard recall points 10,20,...,100. Figures inside the brackets are percentage improvements obtained by use of two and three sources of evidence in the document file, over the results obtained using a single source of evidence.

From Table 3 it can be seen that there is a significant improvement in retrieval effectiveness as we move from the use of single to multiple sources of evidence in the document file,

QUERIES	SOURCES OF EVIDENCE (Documents)						
	Single Evidence			Two Evidences			Combined
	Th	KW	Tx	Th,KW	Th,Tx	KW,Tx	Th,KW,Tx
Th	8.8	-	-	12.1 (+37.8)	14.2 (+61.4)	-	15.1 (+71.1)
KW	-	16.5	-	18.7 (+13.7)	-	26.1 (+58.2)	27.9 (+69.1)
NLQ	-	-	22.3	-	24.0 (+7.9)	24.3 (+9.3)	25.3 (+13.7)

Table 3: Single and multiple sources of evidence in documents

while the number of sources of evidence in the query file remains unaltered. These improvements are obtained due to the probabilistic interpretation of 'relevance' relation between the document evidence and query concepts. This would not be true in the exact-match Boolean retrieval technique, where the number of sources of evidence available is not taken into consideration at all.

While it may be possible to improve the performance of queries represented using thesaurus terms by more careful selection and assignment of these terms to queries, these results show that comparable or better retrieval effectiveness can be obtained by using natural language and user identified query terms and phrases, *with considerably less effort.*

More significantly, irrespective of the performance of a specific index representation in queries, the results show that its presence as an additional source of evidence in the document file can contribute to the improved performance of queries formed using other index representations. This is evident if we compare the figures (Table 3) for 'KW' and 'NLQ' queries on 'KW,Tx' and 'Th,KW,Tx' document index representations. While the evidence provided by thesaurus terms by themselves is quite weak, their presence in the documents improved the performance of KW and NLQ queries.

Hypothesis 2 : Significant improvements in retrieval effectiveness can be obtained by combining results from multiple index representations in queries.

We evaluated the first hypothesis by using different combinations of automatic and manual index representations as sources of evidence in the document file and studied their retrieval performance on queries expressed in a single index representation. To test the second hypothesis, we combined manual and automatic index representations in queries and studied their retrieval performance on the same document file. We generated four query files, using the following combinations of index representations :

1. Thesaurus and keyword queries (Th,KW),

2. Thesaurus and natural language queries (Th,NLQ),

3. Keyword and natural language queries (KW,NLQ), and

4. Combined query file of Th,KW, and NLQ (Th,KW,NLQ).

We used the same query files (i.e., Th, KW and NLQ) that were used in the first set of experiments to generate these combinations. Within each file, different index representations of a query were combined using the #sum operator. For example, the format of the combination of a query expressed as NLQ and using the thesaurus terms would be #sum(#sum(Th) #sum(NLQ)).

These four query files were processed on the combined document index file of 'Th,KW,Tx' and the results evaluated with the standard relevance judgements. We used the combined document index file for these tests as this had produced the best results in the first set of experiments. The results are given in Tables 4,5 and 6. In each of these tables, the figures in the first column are for the queries expressed in a single index representation, the figures in second and third columns are for the queries expressed as a combination of two index representations and the figures in the fourth column are for the combination of three index representations.

As general observations, it may be seen from the results shown in Tables 4, 5 and 6 that 1) adding automatic index representations (i.e., natural language query NLQ) to manual index representations (i.e., keywords and thesaurus terms) in queries significantly improves the performance of these representations, 2) the performance of controlled vocabulary terms in queries can be significantly improved by combining them with the natural language query or keywords selected from the query, and 3) queries formulated using keywords, i.e., terms and phrases selected (manually in these experiments) from the natural language queries, tend to outperform the other two index representations, confirming the importance of phrases in query representations (Croft et al., 1991).

While the retrieval effectiveness of either the automatic index or keyword representation in queries could be improved by their combined representation, the same was not true when they were combined with thesaurus terms (see Tables 5 and 6). Addition of thesaurus terms to automatic index representations (NLQ) improved precision at middle and high recall points, while lowering precision at low recall points. Their addition to keywords lowered precision at all recall levels except the highest. But when all three index representations

	Precision (% change) – 50 queries			
Recall	Th	Th,NLQ	Th,KW	Th,KW,NLQ
10	34.8	49.6 (+42.2)	52.2 (+49.7)	58.7 (+68.5)
20	28.1	42.8 (+52.5)	45.9 (+63.4)	50.1 (+78.6)
30	23.1	35.3 (+52.7)	40.4 (+74.9)	43.4 (+87.6)
40	18.5	29.0 (+56.7)	31.9 (+72.4)	35.3 (+90.3)
50	14.4	22.5 (+56.0)	25.9 (+79.7)	28.3 (+95.9)
60	11.3	18.9 (+67.0)	20.7 (+82.9)	23.8(+111.0)
70	9.0	13.8 (+52.5)	16.1 (+77.8)	18.2(+101.7)
80	7.1	10.6 (+48.1)	11.6 (+63.4)	13.4 (+88.4)
90	3.6	6.1 (+71.2)	6.5 (+81.9)	8.2(+130.3)
100	0.8	2.4(+186.3)	3.0(+254.5)	3.3(+287.0)
average	15.1	23.1 (+53.1)	25.4 (+68.5)	28.3 (+87.4)

Table 4: Combining thesaurus terms with keyword and automatic index representations in queries

were combined in the queries, thesaurus terms helped in improving precision at middle and high recall levels, while lowering precision at the top two recall levels. To see why this was happening, we looked at the probabilities (belief estimates) produced by these three representations and noticed that the probabilities produced by thesaurus terms were much higher than that produced by keywords and NLQ. These higher probabilities seem to be produced due to the low collection frequencies of these terms in the test collection resulting in high inverse document frequencies. Consequently, when the rankings are combined, documents retrieved by thesaurus terms, which include both relevant and non relevant documents, tend to dominate the relevant documents retrieved by other index representations. We reformulated the combined index representation query as a weighted sum query (#wsum operator) and ran a series of experiments lowering the weight of thesaurus queries. The best performance was achieved when the thesaurus term queries were scaled by a factor of 0.3. The results are given in Table 7. Similar results have been reported with respect to ACM CR classification categories in the CACM test collection (Turtle, 1990).

Hypothesis 3 : Significant improvements in retrieval effectiveness can be obtained by combining results from multiple query strategies.

The query strategies we investigate here are the natural language queries (NLQ), Boolean queries (BOOL) and weighted term queries (WTERM). The rationale for using these search strategies for evaluating this hypothesis is that most operational retrieval systems use Boolean queries and these are usually constructed from the natural language description of the user's

	Precision (% change) – 50 queries			
Recall	KW	Th,KW	KW,NLQ	Th,KW,NLQ
10	64.3	52.2 (−18.8)	65.2 (+1.5)	58.7 (−8.6)
20	53.6	45.9 (−14.4)	54.7 (+2.1)	50.1 (−6.5)
30	41.7	40.4 (−3.1)	45.5 (+9.2)	43.4 (+4.0)
40	31.9	31.9 (+0.1)	34.1 (+6.8)	35.3 (+10.6)
50	27.1	25.9 (−4.3)	27.8 (+2.7)	28.3 (+4.3)
60	22.3	20.7 (−7.2)	23.3 (+4.5)	23.8 (+7.0)
70	16.9	16.1 (−5.2)	17.8 (+5.2)	18.2 (+7.6)
80	12.1	11.6 (−3.9)	13.2 (+8.8)	13.4 (+10.8)
90	7.1	6.5 (−9.4)	7.9 (+10.9)	8.2 (+14.7)
100	1.6	3.0 (+82.4)	2.0 (+24.7)	3.3 (+99.2)
average	27.9	25.4 (−8.8)	29.2 (+4.6)	28.3 (+1.4)

Table 5: Combining keywords with thesaurus and automatic index representations in queries

information needs. By way of additional information that can facilitate Boolean query formulation, many of these systems also collect from the user a list of terms to be used for searching, and the importance they attach to these terms. Given this, we felt it would be interesting to find out the improvements that can be obtained by combining these search strategies.

We constructed Boolean queries using keywords and combined these as separate queries with NLQ queries using the #sum operator. Boolean, NLQ and their combined representations were then processed separately on the combined document index file. The results are given in Table 8.

In earlier experiments with the inference net model much better improvements have been reported for the CACM collection (Turtle & Croft, 1991b). The improvements depend on the quality of the Boolean queries. It may be possible to obtain significant improvements by more careful construction of Boolean queries incorporating domain specific knowledge (e.g., thesaurus terms), synonyms, etc. Another interesting way of looking at these results is that, in the absence of good Boolean queries, equally impressive results can be obtained by probabilistic processing of NLQ queries alone.

Here, the interpretation of Boolean queries is probabilistic, which has been shown to perform much better than exact-match interpretation in earlier experiments (Turtle, 1990). In Table 9 we show the difference between the exact-match (E-BOOL) and probabilistic interpretation of Boolean queries (P-BOOL) used in these experiments, for the INSPEC test

	Precision (% change) – 50 queries			
Recall	NLQ	Th,NLQ	KW,NLQ	Th,KW,NLQ
10	63.8	49.6 (−22.3)	65.2 (+2.3)	58.7 (−7.9)
20	50.2	42.8 (−14.8)	54.7 (+8.9)	50.1 (−0.2)
30	38.2	35.3 (−7.5)	45.5 (+19.4)	43.4 (+13.7)
40	28.6	29.0 (+1.4)	34.1 (+19.0)	35.3 (+23.2)
50	23.5	22.5 (−4.4)	27.8 (+18.2)	28.3 (+20.0)
60	17.9	18.9 (+5.4)	23.3 (+30.0)	23.8 (+33.2)
70	13.3	13.8 (+3.8)	17.8 (+34.3)	18.2 (+37.3)
80	9.9	10.6 (+6.6)	13.2 (+33.1)	13.4 (+35.6)
90	5.9	6.1 (+2.6)	7.9 (+33.4)	8.2 (+38.0)
100	1.8	2.4 (+35.8)	2.0 (+14.9)	3.3 (+83.6)
average	25.3	23.1 (−8.8)	29.2 (+15.2)	28.3 (+11.7)

Table 6: Combining automatic index terms with thesaurus and keywords in queries

collection.

We constructed a weighted sum query (#wsum) of keywords by assigning weights to individual keywords on a scale of two importance levels - very important and less important, based on a careful analysis of the natural language queries. This weighted sum query was combined with the NLQ as a separate query using the probabilistic sum operator. The results of processing these three query files (WTERM, NLQ and the combined query strategy file) on the combined document index file is given in Table 10. It can be seen that significant improvements can be obtained by combining these query strategies.

While constructing the weighted term queries, we also considered whether different scales of term weights made any difference to search results. In addition to assigning weights on a scale of two importance levels (very important and less important), we also separately assigned three level weights - very important, moderately important and less important. The results are given in Table 11. It appears that a scale of two weights perform as well as a scale of three weights. While more experiments are required in this direction, we believe these results have implications for query acquisition - manual or automatic.

5 Conclusion

Based on the results in Section 4, we can accept the first two hypotheses that, by treating manual and automatic index representations in queries and documents as sources of evidence, significant improvements in retrieval effectiveness can be obtained by combining these sources

	Precision (% change) – 50 queries	
Recall	Th,KW,NLQ	Th,KW,NLQ (Th 0.3)
10	58.7	66.5 (+13.3)
20	50.1	55.1 (+10.0)
30	43.4	47.8 (+10.3)
40	35.3	38.1 (+8.1)
50	28.3	32.2 (+14.0)
60	23.8	25.0 (+5.1)
70	18.2	19.8 (+8.5)
80	13.4	14.0 (+4.5)
90	8.2	8.7 (+6.4)
100	3.3	2.6 (−19.5)
average	28.3	31.0 (+9.7)

Table 7: Reducing the weight of thesaurus terms

of evidence in the inference net probabilistic retrieval model. The results also support the third hypothesis that by combining different query strategies we can obtain results that are much better compared to the use of these strategies on their own. The results reported in this paper for the combined index and query representations are much better than the results that have been reported in earlier experiments using the INSPEC test collection (Salton & Buckley, 1988; Fox & Koll, 1988). Furthermore the results also show that, while automatic index representations and natural language queries can be combined with manual index representations and Boolean queries to obtain improved performance, they can be used on their own to obtain comparable retrieval performance in situations where manual index representations and Boolean queries are not available or cannot be produced cost effectively.

We believe these results have practical implications for operational information retrieval systems in the sense that by adapting probabilistic retrieval techniques they could more fully exploit the different 'clues' that exist in documents and natural language descriptions of user information needs. The perceived computational complexity of best-match retrieval models have been an hindrance for their use in large scale information services until recently (Rajashekar, 1988). Given the processing capabilities of present day workstations, the availability of inexpensive storage options, and the demonstrated efficient implementation of these models (Turtle & Croft, 1991a; Harman & Candela, 1991), the situation is ripe for wider use of these techniques.

	Precision (% change) – 50 queries		
Recall	BOOL	NLQ	Combined
10	55.6	63.8 (+14.8)	59.7 (+7.5)
20	44.4	50.2 (+13.2)	47.9 (+7.9)
30	37.5	38.2 (+1.7)	40.6 (+8.1)
40	29.8	28.6 (−3.9)	32.3 (+8.4)
50	25.9	23.5 (−9.2)	27.2 (+5.0)
60	21.3	17.9 (−15.9)	22.6 (+6.3)
70	17.1	13.3 (−22.4)	18.1 (+6.1)
80	10.6	9.9 (−6.7)	13.9 (+30.6)
90	6.4	5.9 (−6.9)	8.4 (+31.2)
100	1.5	1.8 (+15.1)	1.9 (+23.4)
average	25.0	25.3 (+1.2)	27.3 (+9.0)

Table 8: Combining Boolean and NLQ query strategies

References

Belkin, N. J. & Croft, W. B. (1987). Retrieval techniques. In Williams, M. E. (Ed.), *Annual Review of Information Science and Technology*, chapterc4, pages 109–145. Elsevier Science Publishers.

Bookstein, A. (1985). Probability and fuzzy-set applications to information retrieval. In Williams, M. E. (Ed.), *Annual Review of Information Science and Technology*, pages 117–151. Knowledge Industries Publications, Inc.

Callan, J. P., Croft, W., & Harding, S. (1992). The INQUERY retrieval system. In Tjoa, A. & Ramos, I. (Eds.), *Database and Expert Systems Applications : Proceedings of the International Conference in Valencia, Spain, 1992*. Springer-Verlag.

Cooper, W. S. (1988). Getting beyond Boole. *Information Processing and Management*, 24(3):243–248.

Croft, W. B. (1987). Approaches to intelligent information retrieval. *Information Processing and Management*, 23(4):249–254.

Croft, W. B. & Harper, D. J. (1979). Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35:285–295.

	Precision (% change) – 50 queries		
Recall	E-BOOL	P-BOOL	
10	39.1	55.6	(+42.3)
20	29.4	44.4	(+51.0)
30	23.7	37.5	(+58.3)
40	15.9	29.8	(+87.1)
50	11.1	25.9	(+133.1)
60	8.4	21.3	(+154.6)
70	5.5	17.1	(+209.9)
80	2.2	10.6	(+378.6)
90	0.9	6.4	(+605.9)
100	0.4	1.5	(+278.3)
average	13.7	25.0	(+83.1)

Table 9: Exact match and probabilistic interpretation of Boolean queries

Croft, W. B., Turtle, H., & Lewis, D. (1991). The use of phrases and structured queries in information retrieval. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 32–45.

DARPA (1990). Tipster information package. Information package related to Defense Advanced Research Project Agency BAA 90-16.

Fox, E. A. & Koll, M. B. (1988). Practical enhanced Boolean retrieval: Experiences with the SMART and SIRE systems. *Information Processing and Management*, 24(3):257–267.

Fox, E. A., Nunn, G. L., & Lee, W. C. (1988). Coefficients for combining concept classes in a collection. In *Proceedings of the Eleventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 291–308, New York, NY. ACM.

Harman, D. & Candela, S. (1991). Retrieving records from a gigabyte of text on a minicomputer using statistical ranking. *Journal of the American Society for Information Science*, 41(8):581–589.

Katzer, J., McGill, M. J., Tessier, J. A., Frakes, W., & DasGupta, P. (1982). A study of the overlap among document representations. *Information Technology: Research and Development*, 1:261–274.

	Precision (% change) – 50 queries		
Recall	WTERM	NLQ	Combined
10	64.3	63.8 (−0.8)	68.3 (+6.2)
20	52.1	50.2 (−3.6)	57.0 (+9.4)
30	44.0	38.2 (−13.4)	47.7 (+8.4)
40	34.1	28.6 (−16.2)	37.0 (+8.4)
50	28.7	23.5 (−18.1)	30.5 (+6.2)
60	23.0	17.9 (−22.1)	24.2 (+5.1)
70	17.5	13.3 (−24.2)	18.4 (+5.3)
80	13.5	9.9 (−26.5)	14.6 (+8.7)
90	7.5	5.9 (−21.0)	7.9 (+6.0)
100	1.7	1.8 (+6.2)	2.0 (+20.5)
average	28.7	25.3 (−11.7)	30.8 (+7.4)

Table 10: Combining weighted term and NLQ query strategies

Maron, M. E. (1988). Probabilistic design principles for conventional and full-text retrieval systems. *Information Processing and Management*, 24(3):249–255.

McGill, M., Koll, M., & Noreault, T. (1979). An evaluation of factors affecting document ranking by information retrieval systems. Technical report, Syracuse University, School of Information Studies.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers.

Rajashekar, T. B. (1988). Improving SDI systems : Implications of new retrieval models. *Library Science with a Slant to Documentation*, 25:44–62.

Robertson, S. E. (1977). The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304.

Salton, G. & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(3):513–524.

Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Sparck Jones, K. & Bates, R. G. (1977). Research on automatic indexing 1974–1976. Technical report, Computer Laboratory, University of Cambridge.

	Precision (% change) – 50 queries		
Recall	KW(No weights)	KW (3 weights)	KW(2 weights)
10	64.3	64.2 (−0.1)	64.3 (+0.1)
20	53.6	51.8 (−3.3)	52.1 (−2.7)
30	41.7	42.8 (+2.7)	44.0 (+5.6)
40	31.9	34.2 (+7.1)	34.1 (+7.0)
50	27.1	28.4 (+5.0)	28.7 (+6.1)
60	22.3	22.7 (+1.8)	23.0 (+3.2)
70	16.9	17.1 (+1.3)	17.5 (+3.5)
80	12.1	13.4 (+10.6)	13.5 (+11.2)
90	7.1	7.4 (+4.4)	7.5 (+5.2)
100	1.6	1.7 (+2.4)	1.7 (+2.1)
average	27.9	28.4 (+1.9)	28.7 (+2.8)

Table 11: Two and three levels of term importance

Sparck Jones, K. & van Rijsbergen, C. J. (1976). Information retrieval test collections. *Journal of Documentation*, 32(1):59–75.

Turtle, H. & Croft, W. B. (1990). Inference networks for document retrieval. In Vidick, J.-L. (Ed.), *Proceedings of the 13th International Conference on Research and Development in Information Retrieval*, pages 1–24. ACM.

Turtle, H. & Croft, W. B. (1991a). Efficient probabilistic inference for text retrieval. In *Proceedings RIAO 3*, pages 644–661.

Turtle, H. & Croft, W. B. (1991b). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187–222.

Turtle, H. R. (1990). *Inference Networks for Document Retrieval.* PhD thesis, University of Massachusetts at Amherst.

Wagers, R. (1992). DowQuest and Dow Jones Text-Search : Which works best and when? *Online*, 16(6):35–41.