An Evaluation of Query Processing Strategies Using the TIPSTER Collection

James P. Callan and **W. Bruce Croft**
Computer Science Department
University of Massachusetts, Amherst, MA 01003, USA
callan@cs.umass.edu, croft@cs.umass.edu

Abstract

The TIPSTER collection is unusual because of both its size and detail. In particular, it describes a set of information needs, as opposed to traditional queries. These detailed representations of information need are an opportunity for research on different methods of formulating queries. This paper describes several methods of constructing queries for the INQUERY information retrieval system, and then evaluates those methods on the TIPSTER document collection. Both AdHoc and Routing query processing methods are evaluated.

1 Introduction

One approach to improving the effectiveness of an information retrieval (IR) system is to use sophisticated methods of gathering and representing information from a user. Techniques include automatic or interactive introduction of synonyms [Har88], forms-based interfaces [CD90], automatic recognition of phrases [CTL91], and relevance feedback [SB90]. All of these techniques have shown promise on standard test collections, but it was not clear how they would scale up to much larger and more heterogeneous document collections.

A large and heterogeneous document collection for IR research became available recently as a result of the DARPA/SISTO TIPSTER project [Har92a]. The first two volumes of the TIPSTER collection contain the full text of about 750,000 newspaper articles, newswire articles, magazine articles, Federal Register announcements, and Department of Energy technical abstracts. These two volumes occupy about two gigabytes of disk space. Instead of standard query sets, TIPSTER information needs are described by frame-like data structures called *topics* (Figure 1). There are two sets of fifty topics.[1] A topic consists of header fields, and seven fields describing aspects of the information need: Domain, Title, Description, Narrative, Concepts, Factors, and Definitions. Each field provides a view of the information need that is related to, but often distinct from, the views provided by the other fields. In particular, the Narrative field is a natural language description of the conditions that make a document relevant to a topic. Relevance judgements for the TIPSTER collection were obtained by having trained analysts evaluate the top documents retrieved for each topic by a variety of different information retrieval systems. In our experiments, we used the relevance judgements from the first Text REtrieval and Evaluation (TREC1) conference [Har92b]. The relevance judgements should be considered incomplete, because most documents were not evaluated for relevance to any topic.

The TIPSTER collection differs in many characteristics from the standard test collections available previously. One difference is the set of TIPSTER topics, which each contain varying representations of an information need. These different representations encourage research on how best to acquire and represent an information need. For example, one can experiment with fill-in-the-blank, forms-like, interfaces by creating queries with text from the Concept(s) and Factors fields. One can also experiment with natural language interfaces by using text from the Narrative and Definitions fields. The research results obtained are suggestive of what might work well in an interactive interface with a human user.

Previous research with other collections suggested that combining different representations of an information need can yield an improvement in both recall and precision [KMT+82; CD90; Tur91]. However, it has been difficult to do systematic research on this subject because different representations of information needs have not been available generally.

The TIPSTER project also distinguishes among two different types of queries: AdHoc and Routing [Har92a]. *AdHoc* queries are intended for a single use to satisfy an immediate need for information. One

[1] A third volume of documents and a third set of fifty topics are planned.

<top>
<head> Tipster Topic Description
<num> Number: 004
<dom> Domain: International Finance
<Title> Topic: Debt Rescheduling
<desc> Description: Document will discuss a current debt rescheduling agreement between a developing country and one or more of its creditor(s).
<narr> Narrative:
A relevant document will discuss a current debt rescheduling agreement reached, proposed, or being negotiated between a debtor developing country and one or more of its creditors, commercial and/or official. It will identify the debtor country and the creditor(s), the repayment time period requested or granted, the monetary amount requested or covered by the accord, and the interest rate, proposed or set.
<con> Concept(s):
1. rescheduling agreement, accord, settlement, pact
2. bank debt, commercial debt, foreign debt, trade debt, medium-term debt, long-term debt
3. negotiations, debt talks
4. creditor banks, creditor countries/governments, Paris Club
5. debtor countries, developing countries
6. debt package
7. debt repayments
8. restructuring, rescheduling existing loans
9. lower interest-rate margin, easier terms, more lenient terms
<fac> Factor(s):
<nat> Nationality: Developing country
<time> Time: Current
</fac>
<def> Definition(s):
Debt Rescheduling - Agreement between creditors and debtor to provide debt relief by altering the original payment terms of an existing debt. This is most often accomplished by lengthening the original schedule for principal and interest payments, and deferring interest payments. Done most publicly by developing countries and their bankers, but often less publicly by other willing creditors and debtors, e.g., governments, banks and companies. Much in vogue
in the early 1980s, the road to rescheduling for countries in crisis runs as follows: when a country borrows so much that its lenders grow nervous, the banks start lending for shorter and shorter maturities. Eventually the country, though still paying interest on its debt, is unable to make payments on the principal. The country is then forced to request a rescheduling, which means that it is able to escape its immediate repayment commitments by converting short-term loans into longer-term ones. A country wishing to reschedule its official debt talks to the Paris Club. A country wishing to reschedule its commercial debt talks to its biggest bankers.
</top>

Figure 1: A TIPSTER topic.

might expect to invest only a moderate effort in creating an AdHoc query, because the query is used once and then discarded. *Routing* queries (sometimes called SDI queries) are intended for longer-term use. A routing query can be viewed as a profile of interest, or as a filter on a steady stream of incoming documents. One might expect to invest greater effort in creating routing queries, because the effort is amortized over many retrievals.

This paper describes a set of experiments on the effectiveness of different methods of query creation. The experiments explore what to include in a query, how to represent it, and how to combine different representations of the same information need. Section 2 describes the INQUERY information retrieval system, with which experiments were conducted. Section 3 describes techniques for creating 'AdHoc' queries. Section 4 describes techniques for creating 'Routing' queries. Section 5 summarizes the results, and concludes.

2 The INQUERY Information Retrieval System

INQUERY is a probabilistic information retrieval system based upon a Bayesian inference network model [TC91; Tur91]. Documents are indexed by the word stems and numbers that occur in the text. Stopwords are discarded. Documents are also indexed automatically by a small number of *features* that provide a controlled indexing vocabulary [CCH92].[2] For example, when a document refers to a company by name, the document is indexed by both the company name (words in the text) and the feature #company.[3] INQUERY includes company [Rau91], country, U.S. city,

[2] We define a *feature* to be any generalization of words in a document. We have sometimes called these generalizations *concepts* (e.g., [CCH92]).

[3] INQUERY distinguishes among features and words by prefixing features with the '#' character.

number and date [Mau89], and person name recognizers. The set used for a particular collection can be controlled easily, and new, domain-specific recognizers can be incorporated easily [CCH92]. It remains an open question how to determine the 'right' mix of feature recognizers for a document collection.

The query language contains about a dozen operators [TC91; Tur91]. *Feature operators* match features that were recognized when the document was parsed. For example, the #company operator matches the #company feature. *Proximity operators* require their arguments to occur either in order, within some distance of each other, or within some window. *Belief operators* provide different methods of combining evidence. Examples include using the maximum, sum, or weighted sum of a set of beliefs. INQUERY also has a synonym operator and probabilistic versions of Boolean And, Or and Not operators.

The query processor provides several transformations that aid in converting queries from natural language into the query language. These *query transformations* include recognition of stop phrases, negation, phrases, and proper names, as well as introduction of synonyms and controlled vocabulary terms (feature operators). Each is discussed in more detail below. INQUERY allows the query transformations to be combined to form different methods of query processing.

2.1 Stop Phrases

Stop phrases are sequences of words that are discarded automatically because they provide no information about the information need. For example, the phrases "document must discuss" and "find a document" are both stop phrases. The list of stop phrases was created manually after examination of 50 TIPSTER topics. It also includes a small number of phrases that do not occur in the topics, but that would be likely to occur in an interactive system.

2.2 Phrase Recognition

Phrases are recognized in the query by applying a stochastic part of speech tagger [Chu88], and then using rules to identify noun phrases. For example, "monthly short interest" is transformed into "#phrase (monthly short interest)". Experiments showed that simple noun phrases worked best, because longer, more complex, noun phrases were less likely to match documents in the collection.

2.3 Negation

Negation is recognized by looking for the word 'not' in the query, and then negating the object (word, proper name, phrase or query language operator) that immediately follows. For example, "not #phrase (monthly short interest)" is transformed into "#not (#phrase (monthly short interest))". This strategy is too simple to do justice to negation in the English language, but it does provide some improvement, at low recall, over ignoring negation altogether. Table 1 illustrates the effect of negation on the five queries in which it was recognized automatically.

Table 1: The effect of negation.

Recall	Precision (5 queries)		
	Not	— No Not —	
25	60.9	60.6	(−0.6)
50	49.5	49.6	(+0.1)
75	32.7	33.1	(+1.5)
avg	47.7	47.8	(+0.1)

2.4 Proper Names

Proper names are recognized by assuming that a sequence of capitalized words is a proper name. Commas and other punctuation are also assumed to delimit proper names. This strategy is too simple to find all proper names (e.g. John von Neumann), but it works often. (Proper names that escape detection are generally recognized as noun phrases, as described above.) A proximity operator is used to match recognized proper names against documents in the collection. The proximity operator requires that its arguments occur in a document, in order, with an interword distance of three or less. This permits the query "George Bush" to match "George Herbert Walker Bush" in a document.

2.5 Synonyms

The use of synonyms is currently limited in INQUERY. We focused on a small set of words that occur in the Factors field of TIPSTER topics, because those concepts are supposed to be particularly important in determining the relevance of a document. We have experimented with replacing "Europe" with a list of European countries, and with replacing "developing country", "third world", and "LDC" with a list of developing countries. We have also tried replacing these words with a negated list of their opposites, for example replacing "Europe" with "not USA, Canada, Mexico, ...". None of these changes produced an average improvement, although some yielded small improvements in precision at low recall.

2.6 Feature Operators

Certain words in the query cause the introduction of feature operators that match any occurrence of a particular feature in a document. For example, if the word 'company' occurs in a query, it is replaced by the operator #company, which matches any company. Similar expansions occur for references to foreign countries, US cities, and the United States. In general, the #company operator was the most effective for TIPSTER queries. Replacing "United States" and its variations with the #usa feature generally hurt

Table 2: The effect of replacing the query word "location" with the concepts #us-city and #foreigncountry.

Recall	Precision (8 queries)				
	NoCity	— City —		– City+ForeignCountry –	
25	45.8	46.7	(+2.0)	46.8	(+2.3)
50	30.3	30.4	(+0.2)	30.7	(+1.2)
75	15.0	14.9	(−1.2)	15.2	(+1.4)
avg	30.4	30.6	(+0.9)	30.9	(+1.8)

performance slightly. We believe that this occurred because most documents in this database are in some way about the United States. Replacing "United States" with "not #foreigncountry" was more effective. Replacing the word "location" with the concepts #us-city and #foreigncountry yielded a small average improvement on the 8 queries in which the word occurs (Table 2).

3 Techniques for Creating AdHoc Queries

We adopted three different approaches to creating AdHoc queries. The first approach used only the contents of the Description field of TIPSTER topics. This was useful for exploring how the system behaves with the very short queries. The second approach used the contents of the Description, Title, Narrative, Concept(s) and Factor(s) fields. This was useful for exploring how a system might behave with an elaborate user interface or very sophisticated query processing. The third approach was automatic query creation followed by simple manual modifications, to simulate simple user interaction with the query processing. Each approach is described below in more detail.

3.1 Simple Queries

The *description-only* approach to query processing treated the Description field of a TIPSTER topic as if it were the only user input. A query was constructed, automatically, by employing all of the query processing transformations described above (phrase identification, stop phrase removal, synonym expansion, proper name recognition, etc). The remaining words and operators were enclosed in a weighted sum operator, with weights determined by frequency in the query.

Results are summarized in Table 3. Provided are both a traditional recall/precision table and a table showing precision in the top n documents retrieved, for 5 values of n (5, 15, 30, 100, 200). The recall/precision table is provided because it measures the ability of the IR system to locate all of the documents known to be relevant to each query. The precision in the top n documents retrieved is provided both for comparison to other TREC1 results, and because it is a better measure of what a person using the system would see. The results were obtained by creating queries for TIPSTER topics 51-100, and using them to retrieve documents from Volume 1 of the TIPSTER data. The

Table 3: The performance of simple 'description-only' queries.

Recall	Precision
0	62.0
10	36.3
20	29.7
30	25.3
40	22.2
50	18.9
60	15.6
70	12.2
80	9.6
90	6.0
100	0.9
avg	21.7

Recall (#Docs)	Precision
5	0.460
15	0.380
30	0.333
100	0.246
200	0.200

TREC1 relevance judgements were used to determine relevance.

The results with the 'description-only' queries are surprisingly good, given their brevity, the size of the document collection, and the difficulty of some of the topics. The set of documents retrieved by the description-only queries is quite different from the sets retrieved by longer queries. Some documents that were ranked highly by the description-only queries had no relevance judgements, so it is unclear whether the documents were relevant (but not judged), or not relevant. A similar phenomenon has been identified with short Boolean queries [BCCC93].

3.2 Multiple Sources of Information

The *multiple-field* approach to query processing applied different types of processing to different fields. Experiments were conducted with a variety of processing combinations. Results for six of these combinations are reported below. In general, most of the query processing transformations described above were applied to each field. The exceptions were the Narrative field, and the Concept(s) field. The text in the Narrative field was usually a very abstract discussion of the criteria for document relevance. Such a discussion is not well-suited to a system like INQUERY, which relies on matching words from the query to words in the document. In contrast, the Concepts field was highly structured. Phrases and proper names were always delimited by commas or periods, making syntactic recognition of phrases unnecessary.

Experiments were conducted with six methods of

Table 4: A comparison of six automatic methods of constructing AdHoc queries.

Recall	Precision (50 queries)									
	Q-1	—- Q-3 —-		—- Q-4 —-		—- Q-6 —-		—- Q-F —-		—- Q-7 —-	
0	83.9	83.2	(−0.8)	78.8	(−6.1)	86.2	(+2.7)	83.0	(−1.1)	84.7	(+1.0)
10	60.5	59.0	(−2.5)	57.3	(−5.2)	61.6	(+1.9)	60.9	(+0.7)	61.8	(+2.2)
20	52.7	49.9	(−5.4)	49.0	(−7.1)	52.3	(−0.8)	53.1	(+0.6)	53.5	(+1.4)
30	46.6	45.0	(−3.6)	44.0	(−5.6)	46.4	(−0.4)	48.2	(+3.3)	47.6	(+2.0)
40	40.5	40.0	(−1.2)	39.6	(−2.3)	40.5	(+0.0)	41.9	(+3.5)	42.2	(+4.2)
50	35.0	35.4	(+1.2)	34.6	(−1.0)	35.9	(+2.6)	36.5	(+4.3)	36.8	(+5.1)
60	30.5	30.4	(−0.5)	29.0	(−4.9)	30.3	(−0.8)	31.4	(+3.1)	31.4	(+3.0)
70	25.4	25.5	(+0.5)	24.3	(−4.5)	24.1	(−5.0)	26.3	(+3.5)	26.1	(+2.9)
80	19.9	19.7	(−0.6)	18.6	(−6.6)	17.5	(11.9)	20.0	(+0.7)	19.7	(−0.9)
90	12.1	12.4	(+2.3)	11.3	(−6.9)	11.0	(−8.9)	13.0	(+7.6)	12.2	(+0.8)
100	2.5	2.6	(+5.4)	2.5	(−0.2)	2.0	(18.4)	2.5	(+3.1)	2.4	(−4.1)
avg	37.2	36.6	(−1.6)	35.4	(−5.0)	37.1	(−0.4)	37.9	(+1.8)	38.0	(+2.1)

Recall	Precision (50 queries)									
(#Docs)	Q-1	—- Q-3 —-		—- Q-4 —-		—- Q-6 —-		—- Q-F —-		—- Q-7 —-	
5	64.8	62.8	(−3.1)	61.2	(−5.6)	64.8	(+0.0)	67.2	(+3.7)	67.2	(+3.7)
15	59.2	55.6	(−6.1)	54.7	(−7.6)	59.6	(+0.7)	59.7	(+0.9)	60.7	(+2.5)
30	54.1	53.3	(−1.5)	51.3	(−5.2)	54.5	(+0.7)	55.0	(+1.7)	55.9	(+3.3)
100	42.4	42.2	(−0.5)	41.6	(−1.9)	43.6	(+2.8)	44.0	(+3.8)	43.6	(+2.8)
200	35.6	35.1	(−1.4)	34.5	(−3.1)	36.0	(+1.1)	36.6	(+2.8)	36.4	(+2.3)

query processing described below. The abbreviations in the descriptions refer to the first letter of a field name (i.e., D means the Description field).

Q-1: Created automatically, using T, D, N, C and F fields. Everything except the synonym and concept operators was discarded from the Narrative field.

Q-3: The same as Q-1, except that recognition of phrases and proper names was disabled.

Q-4: The same as Q-1, except that recognition of phrases was applied to the Narrative field.

Q-6: The same as Q-1, except that only the T, C and F fields were used.

Q-F: The same as Q-1, with 5 additional thesaurus words or phrases added automatically to each query.

Q-7: A combination of Q-1 and Q-6.

Q-1 was the first method tested. It became the baseline method against which other methods were compared. The Q-3 method was a 'words only' query used to determine whether phrase and proximity operators were helpful. The Q-4 method was developed to determine whether the simple query processing transformations would be effective on the abstract descriptions in the Narrative field. The Q-6 method narrowed in on the set of fields that appeared most useful. The Q-F method was a preliminary investigation of an approach to automatically discovering thesaurus terms. The Q-7 method investigated whether combining the results of two relatively similar queries could yield an improvement.

The results from the experiments are summarized in Table 4. The results were obtained by creating queries for TIPSTER topics 51-100, and using them to retrieve documents from Volume 1 of the TIPSTER data. The TREC1 relevance judgements were used to determine relevance.

The difference between the performance of methods Q-1 and Q-3 shows that phrases, proper names and proximity operators were useful. This result confirms previous research showing that phrases improved performance [CTL91]. However, most of the improvement occurred at low recall, resulting in a small average improvement (1.6%). Experiments with different phrase operators produced only small (usually negative) changes in recall and precision. The reason for these results is unclear. Although some phrases were more common in the collection than others, we do not believe the phrases themselves were the problem. The INQUERY phrase operators treat as individual words any phrases that they consider to be 'low quality', based upon Mutual Information Measure [CGHH91], frequency, or other statistics. It may be that a better phrase operator would solve the problem, or it may be that phrases are less effective on long queries. (The average length of the Q-1 queries is 43.7 words, counting stop words.)

The results for method Q-4 show that phrases from the Narrative were not helpful. This result is not surprising, given the relatively abstract descriptions in this field. However, it would be wrong to interpret this result as indicating that the Narrative is not useful. The Narrative is a statement of what makes a document relevant to the information need. One challenge for future research is to determine how to make better use of this information.

Discarding the Description and Narrative fields did not hurt performance appreciably. Doing so actually improved precision at low (0% and 10%) recall. This result suggests that the Description field contributes little that is not available in other fields of the topic.

Table 5: A comparison of two semi-automatic methods of constructing AdHoc queries.

Recall	Precision (50 queries)				
	Q-1	—- Q-M —-		—- Q-O —-	
0	83.9	83.8	(−0.2)	93.0	(+10.8)
10	60.5	64.1	(+6.0)	71.6	(+18.3)
20	52.7	55.4	(+5.1)	63.4	(+20.3)
30	46.6	48.6	(+4.3)	54.2	(+16.3)
40	40.5	42.1	(+3.9)	46.8	(+15.5)
50	35.0	36.4	(+4.1)	40.4	(+15.6)
60	30.5	30.9	(+1.5)	34.1	(+11.8)
70	25.4	25.0	(−1.4)	28.4	(+11.6)
80	19.9	18.3	(−7.8)	21.7	(+9.1)
90	12.1	11.8	(−3.0)	13.4	(+10.3)
100	2.5	2.3	(−6.5)	2.4	(−2.5)
avg	37.2	38.1	(+2.3)	42.7	(+14.6)

Recall	Precision (50 queries)				
(#Docs)	Q-1	Q-M		Q-O	
5	64.8	67.2	(+3.7)	76.4	(+17.9)
15	59.2	63.9	(+7.9)	72.4	(+22.3)
30	54.1	57.5	(+6.3)	64.9	(+20.0)
100	42.4	45.5	(+7.3)	49.4	(+16.5)
200	35.6	36.7	(+3.1)	39.2	(+10.1)

Only limited use was made of the Narrative field, so it is not surprising that ignoring it completely would have little effect.

The results for method Q-F show that it is possible to automatically construct a useful thesaurus for a collection, based only upon term associations. The thesaurus words and phrases improved precision at almost all levels of recall. These results, while encouraging, raise many questions. The thesaurus words and phrases were identified automatically by their co-occurrence with query terms in the 1987 Wall Street Journal portion of the document collection. It is not clear whether a useful thesaurus can be constructed from the entire collection or a representative sample. It is also unclear how and how many thesaurus words and phrases to add to the query.

A combination of methods Q-1 and Q-6 produced a 2.1% average improvement over either method alone. This result is confirmation of previous research [KMT+82; CD90; Tur91] in two ways. First, it shows that combining different representations of an information need is helpful. Second, it shows that Q-1 and Q-6, which are similar, retrieve different sets of documents.

The differences in results for these query processing methods are relatively small, for two reasons. First, the differences in the methods themselves were intentionally small, in order to isolate the effects of certain transformations or fields. Second, the queries were all so long that any single change was outweighed by what remained constant.

3.3 Interactive Query Creation

Experiments were also conducted to simulate a more interactive approach to query creation. In these experiments, the system created a query using method Q-1 described above, and then a person[4] was permitted to modify the resulting query. The modifications permitted were restricted to adding words from the Narrative field, deleting words or phrases from the query, and indicating that certain words or phrases should occur near each other within a document. The distance restriction was introduced to simulate paragraph-level retrieval [O'C80].

Table 5 summarizes the results of experiments with two slightly different methods of interactive query processing. The differences between the methods are described below.

Q-M: Manual addition of words or phrases from the Narrative, and manual deletion of words or phrases from the query.

Q-O: The same as Q-M, except that the user could also indicate that certain words or phrases must occur within 50 words of each other.

The difference in results obtained by methods Q-1 and Q-M shows that simple user modifications of automatic query processing can yield improvement. Most of the improvement occurred from Recall levels 10-50%, and that performance degraded thereafter. This behavior would be acceptable in an interactive system, because users are not likely to examine all documents retrieved.

The large improvement obtained with method Q-O suggests that paragraph retrieval, as simulated by the "unordered window" operator, significantly improves effectiveness. This result is encouraging. Our future research will consider how to conduct paragraph-level retrieval without user intervention.

A second set of experiments were conducted to determine the effect of thesaurus terms and phrases on queries that were created automatically and modified manually. In these experiments, five additional thesaurus terms or phrases were added automatically to each query in the Q-1, Q-M and Q-O query sets. The terms and phrases selected were the same as those used in the Q-F query set described above. The resulting query sets were Q-F, Q-MF, and Q-OF. Table 6 summarizes the results of evaluating these query sets on Volume 1 of the TIPSTER data, using the TREC1 relevance judgements.

Manual modification of the Q-F query set yielded a 3.8% average improvement. Inclusion of unordered window operators yielded a 13.2% improvement. These results are largely in agreement with the results from the first set of experiments with semi-automatic query creation. In the first set of experiments, the improvements were 2.3% and 14.6%, respectively. However, closer examination of the results reveals that the thesaurus terms and phrases were most effective in the Q-M query set. The thesaurus terms and phrases improved the Q-M query set from

[4] The second author.

Table 7: A comparison of four methods of constructing routing queries. The methods were evaluated on Volume 2 of the TIPSTER document collection.

Recall	Precision (50 queries)						
	Q-1	—- Q-F —-		—- Q-R —-		—- Q-O —-	
0	77.1	75.2	(−2.4)	78.0	(+1.2)	85.4	(+10.8)
10	55.2	56.1	(+1.7)	58.3	(+5.5)	65.6	(+18.9)
20	48.3	49.0	(+1.4)	50.1	(+3.9)	57.9	(+19.9)
30	41.5	43.0	(+3.4)	43.8	(+5.4)	49.7	(+19.6)
40	36.7	37.7	(+2.8)	37.6	(+2.5)	42.8	(+16.8)
50	32.0	32.9	(+3.0)	32.9	(+2.8)	36.3	(+13.5)
60	27.9	27.9	(+0.3)	27.6	(−0.9)	30.7	(+10.3)
70	22.1	22.9	(+3.5)	23.1	(+4.4)	24.6	(+11.4)
80	17.5	18.0	(+2.8)	18.6	(+6.2)	19.1	(+9.4)
90	12.5	12.8	(+2.7)	12.4	(−0.2)	14.0	(+11.9)
100	2.4	2.7	(+12.1)	3.7	(+51.6)	3.7	(+51.9)
avg	33.9	34.4	(+1.4)	35.1	(+3.5)	39.1	(+15.2)
Recall	Precision (50 queries)						
(#Docs)	Q-1	Q-F		Q-R		Q-O	
5	58.4	58.0	(−0.7)	59.6	(+2.1)	69.6	(+19.2)
15	51.5	53.5	(+3.9)	55.9	(+8.5)	61.1	(+18.6)
30	48.7	50.1	(+2.9)	50.4	(+3.5)	56.6	(+16.2)
100	34.6	35.5	(+2.6)	36.0	(+4.1)	39.2	(+13.3)
200	26.3	26.9	(+2.3)	26.1	(−0.8)	28.5	(+8.4)

Table 6: A comparison of two semi-automatic methods of constructing AdHoc queries, with thesaurus terms added.

Recall	Precision (50 queries)				
	Q-F	—- Q-MF —-		—- Q-OF —-	
0	83.0	86.3	(+4.1)	92.9	(+12.0)
10	60.9	64.0	(+5.1)	70.4	(+15.6)
20	53.1	56.4	(+6.4)	62.0	(+16.9)
30	48.2	50.2	(+4.2)	54.6	(+13.3)
40	41.9	44.0	(+5.0)	47.7	(+13.8)
50	36.5	38.2	(+4.9)	40.7	(+11.7)
60	31.4	32.5	(+3.4)	35.2	(+12.1)
70	26.3	26.5	(+0.7)	29.3	(+11.3)
80	20.0	19.5	(−2.5)	22.2	(+10.9)
90	13.0	12.4	(−5.2)	14.2	(+9.1)
100	2.5	2.4	(−6.7)	2.5	(-2.6)
avg	37.9	39.3	(+3.8)	42.9	(+13.2)
Recall	Precision (50 queries)				
(#Docs)	Q-F	Q-MF		Q-OF	
5	67.2	67.6	(+0.6)	75.2	(+11.9)
15	59.7	63.5	(+6.4)	70.8	(+18.6)
30	55.0	57.8	(+5.1)	64.3	(+16.9)
100	44.0	46.5	(+5.9)	49.4	(+12.3)
200	36.6	37.9	(+3.6)	39.8	(+8.7)

38.1% to 39.3%, which is a 3.2% relative improvement. In contrast, the thesaurus terms and phrases improved the Q-O query set from 42.7 to 42.9, which is a 0.5% improvement. It is unclear what caused this difference. It is possible that the small difference is an artifact of the experimental design. Thesaurus words and phrases were added *after* the query was modified, so they were not used in unordered window operators. The effects of the paragraph-like retrieval may have swamped the contribution of the five thesaurus terms and phrases.

4 Techniques for Creating Routing Queries

Routing queries are queries that are designed for long term use. It is fair to assume that care is exercised in the construction of routing queries, because the time spent in query construction is amortized over many retrievals. Our experiments compared the effectiveness of queries created automatically, interactively, and by relevance feedback. The automatic methods Q-1 and Q-F, and the semi-automatic method Q-O, are described above. Relevance feedback is described below.

Relevance feedback was conducted on the Q-1 query set and the TIPSTER Volume 1 documents, using all of the TREC1 relevance judgements. The hypothesis was that relevance feedback on Volume 1 would produce queries suitable for use on Volume 2. The *rdfidf* method [HC93] was used to select five terms to add to each query. Term weights for all terms were determined by the *rtfidf* method [HC93]. This approach to query creation is called Q-R in this paper.

The results from the experiments are summarized in Table 7. The results were obtained by creating queries for TIPSTER topics 51-100, and using them to retrieve documents from Volume 2 of the TIPSTER data. The TREC1 relevance judgements were used to determine relevance.

These results support the conclusion that the introduction of concepts associated with query terms in one document database can improve the effectiveness of the query in another database. This result is supported by experiments with two different techniques for deciding what to add to the query. One technique identified words and phrases that co-occurred with query terms in a sample of the database, regardless of relevance. The other technique identified words

that were more likely to occur in relevant documents. The former approach requires no user intervention, so it could also be used for AdHoc queries. The latter approach requires user intervention.

The impressive results of the Q-O method demonstrate that paragraph-level retrieval is as effective for Routing queries as it was for AdHoc queries.

5 Conclusion

This paper evaluates several approaches to creating AdHoc and Routing queries for the TIPSTER collection. These approaches were developed with the TIPSTER data in mind, but were not "tuned" for the TIPSTER data. In particular, no relevance judgements were available during the time that the query processing methods were developed.

The results demonstrate both the advantage and difficulty of using as many representations as possible of a user's information need. Queries based upon multiple TIPSTER topic fields yielded better recall and precision than queries based upon a single field. However, experiments with the Narrative field showed that some representations require very sophisticated processing. Careless use of the Narrative is worse than ignoring it completely.

Human interaction with automatic query processing was shown to be helpful. This result can be interpreted as suggesting further improvements to automatic query processing. It may also mean that human intuition remains an important part of query formulation.

Paragraph-level retrieval was shown to be effective in large and heterogeneous collections. This result is not surprising. The TIPSTER data includes many long documents. Some Federal Register documents have so many words that they are retrieved, albeit with low ranking, for virtually every query. Our approximation of paragraph-level retrieval with an "unordered window" operator demonstrated that these spurious matches can be eliminated by restricting attention to paragraph-sized portions of the document. Our future work will concentrate on use of actual paragraphs.

Our results also support the conclusion that indexing automatically with controlled vocabulary terms (*features*), in addition to words in the text, can be effective. We have demonstrated small improvements in precision at all recall levels by careful use of company, foreign country, and US city concepts. Our success rate in selecting useful features to include in the vocabulary is about 50%, so far. The use of synonym query networks for "Europe" and "developing country" did not produce an improvement, nor did use of the USA concept.

The experiments suggest that automatic construction of a thesaurus is possible, using only data about co-occurrence of noun phrases in the collection. The results raise many questions about how best to create the thesaurus, and how best to use the thesaurus words and phrases in a query. However, the thesaurus words and phrases were found to be useful in automatic, semi-automatic and routing experiments. The consistency of these results encourages future research.

The results for the Routing query processing methods confirm that it is possible to create queries with one collection and then use them effectively on another, previously unseen, collection. The collections used in these experiments had many similar characteristics, but differed on some characteristics like average document length. The experiments demonstrated the effectiveness of four different methods for creating routing queries. Automatic and semi-automatic methods performed well. Thesaurus terms from one collection were shown to be effective on another similar collection. Relevance feedback on one collection was shown to produce queries that were useful on another similar collection. It is unclear how well thesaurus words and phrases or relevance feedback would perform if the collections differed more. Paragraph-level retrieval appeared to work as well with routing queries as with AdHoc queries.

This evaluation of query processing also raised many questions. One such question is whether special treatment of phrases is helpful in very long queries. The results of our evaluation suggest that phrases do indeed help, but that the improvement is small. It may be that the large number of words in the query effectively disambiguates documents, making phrases unnecessary. It may also be that a better approach to identifying phrases in queries and documents is required.

A second question is how one can use effectively the contents of the Narrative field. Intuition suggests that this field must be useful, because it describes precisely the requirements for relevance. One possibility, which we are investigating, is that the words from the Narrative should not be used directly, but should instead modify or affect the way in which the other fields are processed.

A final question is whether true paragraph-level retrieval will yield results different from the results obtained with the unordered window operator. One might suspect that results will be better, because the TIPSTER collection contains many "News Summary" documents. However, our intuitions with this collection have often been wrong, so further research is needed.

We are currently carrying out a range of more detailed experiments using the TREC1 relevance judgements. The results from these experiments will allow us to tune the query processing techniques and to make more definite conclusions about their relative effectiveness.

Acknowledgements

This research was supported by the NSF Center for Intelligent Information Retrieval at the University of Massachusetts. We thank Steve Harding, John Broglio and Michelle Lamar for helping with this research.

References

[BCCC93] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. The effect of multiple query representations on information retrieval system performance. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993.

[CCH92] James P. Callan, W. Bruce Croft, and Stephen M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83. Springer-Verlag, 1992.

[CD90] W. B. Croft and R. Das. Experiments with query acquisition and use in document retrieval systems. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 349–368, 1990.

[CGHH91] Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. Using statistics in lexical analysis. In U. Zernick, editor, *Lexical Acquisition: Using Online Resources to Build the Lexicon*. Lawrence Erlbaum, 1991.

[Chu88] Kenneth Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing*, pages 136–143, 1988.

[CTL91] W. B. Croft, H.R. Turtle, and D.D. Lewis. 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, 1991.

[Har88] D. Harman. Towards interactive query expansion. In Y. Chiaramella, editor, *Proceedings of the 11^{th} International Conference on Research and Development in Information Retrieval*, pages 321–332. ACM, June 1988.

[Har92a] D. Harman. The DARPA Tipster project. *SIGIR Forum*, 26(2):26–28, 1992.

[Har92b] D. Harman, editor. *The First Text REtrieval Conference (TREC1)*. National Institute of Standards and Technology Special Publication 200-207, Gaithersburg, MD, 1992.

[HC93] David Haines and W. B. Croft. Relevance feedback and inference networks. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993.

[KMT+82] J. Katzer, M. J. McGill, J. A. Tessier, W. Frakes, and P. DasGupta. A study of the overlap among document representations. *Information Technology: Research and Development*, 1:261–274, 1982.

[Mau89] Michael Loren Mauldin. *Information retrieval by text skimming*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburg, PA, 1989.

[O'C80] John O'Connor. Answer-passage retrieval by text searching. *Journal of the American Society for Information Science*, 31(4):227–239, 1980.

[Rau91] Lisa F. Rau. Extracting company names from text. In *Proceedings of the Sixth IEEE Conference on Artificial Intelligence Applications*, 1991.

[SB90] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *JASIS*, 41:288–297, 1990.

[TC91] Howard Turtle and W. Bruce Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3), July 1991.

[Tur91] Howard Robert Turtle. *Inference networks for document retrieval*. PhD thesis, Department of Computer and Information Science, University of Massachusetts, Amherst, 1991.