Document Retrieval and Routing Using the INQUERY System

John Broglio, James P. Callan, W. Bruce Croft and Daniel W. Nachbar
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, Massachusetts 01003, USA

The INQUERY retrieval and routing system, which is based on the Bayesian inference net retrieval model, has been described in a number of papers [5, 4, 10, 11]. In the TREC experiments this year, a number of new techniques were introduced for both the ad-hoc retrieval and routing runs. In addition, experiments with Spanish retrieval were carried out.

1 Description of Ad-Hoc Experiments

For the ad-hoc retrieval experiments, the major changes to the system were the incorporation of passage retrieval, query expansion using PhraseFinder [8], a new estimation technique for indexing probabilities, and new analysis techniques for the TREC topics.

1.1 Query Processing

The following description of query processing emphasizes the differences to the approach used in the previous evaluations, which is described in [5, 3].

The text in all parts of the TREC topics (topic, description, narrative) were treated in the same way. The University of Massachusetts JTAG tagger [12] was used to identify parts of speech for words in the query. Then sequences of nouns, or sequences of adjectives and following nouns, were selected for the #PHRASE operator. (In the past, we had included prepositional phrases which followed a "bare" headnoun. We eliminated these phrases because we found they were not useful.)

Analysis of phrase statistics has shown that phrases of length greater than two words were invariably too restrictive, especially when subphrases of these would have been useful in retrieval. We therefore used only two-word phrases. Sequences of more than two words were broken down into two-word subsequences. For example: *(crude oil price trend)* $\Rightarrow$ *(crude oil), (oil price), (price trend)*. Using this modification of the original phrase extraction procedures, we were able to eliminate a number of special processing steps, and apply noun-phrase-extraction processing to all sections.

We weighted the sections of the query as follows; title $= 3n$; description $= 1n$, narrative $= 1$; where n is the number of terms in the section. This was accomplished simply by duplicating the title three times, the description once, and by using the #SUM of the narrative terms.

1.2 PhraseFinder

PhraseFinder is a technique for corpus-based query expansion [8]. Phrasal concepts are identified using a part-of-speech tagger, and the text contexts of these concepts (words in close proximity) are used to create concept documents. These documents, in turn, are used to create a concept-based INQUERY database. The query is then processed against this database to produce a ranked list of phrases. Phrases from this list are then added to the original query, with appropriate weighting. In [8], phrases were treated differently depending on whether they contained words that were in the original query ("duplicates") or were entirely new ("non-duplicates"). Previous experiments with TIPSTER/TREC databases and topics, which produced consistent effectiveness improvements, were used to determine the number of phrases and the weighting used.

After query processing created a query q from a topic, the query q was used to retrieve 30 associated phrases from a PhraseFinder database. The retrieved phrases were partioned into a ranked set of phrases already in q, and a ranked set of phrases not in q. Each phrase was weighted according to its rank in the set. "Duplicate" phrases were weighted by $1-(r-1)\cdot\frac{1}{30}$, where r was the phrase's rank. "Non-Duplicate" phrases were weighted by $0.6-(r-1)\cdot\frac{0.6}{30}$. Then the phrases and their weights were added to q to form an expanded query q'.

1.3 Passage Retrieval

In [2], we reported experiments that showed significant improvements in retrieval effectiveness are possible when document rankings based on the entire document text are combined with rankings based on the best passages in the documents. Overlapping passages of a fixed-length yielded better results than passages based on document structure (e.g. paragraphs).

In the ad-hoc document retrieval experiments, the expanded query (q' above) was applied to both documents and passages, as described in [2]. The final form of the query, q'', was

#wsum (1.0 2.0 #passage200 (q'') 1.0 (q''))

based upon the results reported in [2].[1]

1.4 Manual Modifications

These queries were generated by simulating some of the modifications a user might be expected to make to an initial query in an interactive environment. The starting point for this experiment was a version of the automatically produced queries (INQ101). These queries included phrases added automatically using the PhraseFinder as described above. Changes to these initial queries were limited to the deletion of spurious (in the user's opinion) words and phrases, modification of weights based on perceived relative importance, and adding proximity restrictions such as are often used in Boolean systems. The following is an example of the type of query produced.

[1]The first 1.0 is a scaling parameter that is used rarely. It was always 1.0 in our experiments.

```
! Document-level query
#WSUM (1.0 1.0  #SUM( tobacco #company  advertising
                      #PHRASE( young people) tobacco target young
                      #PHRASE( tobacco industry) young youth children
                      teenagers adolescents #phrase(high school students)
                      #phrase(college students) )
           3.0 #PHRASE( tobacco industry)  1.0 aims
           1.0 advertising   1.0 young
           3.0  Tobacco  3.0 company  3.0 advertising   1.0 young
           3.0 #uw100( tobacco advertising #syn(young youth teenagers
                       adolescents students children))
   0.3 #3 ( tobacco industry ) 0.3 #3 ( tobacco company )
   0.3 #3 ( tobacco product ) 0.3 #3 ( ad campaign )
   0.3 minor 0.3 snuff 0.3 #3 ( tv commercial )
   0.3 #3 ( advertising campaign ) 0.3 #3 ( marketing executive )
   0.3 teenager 0.3 glamour 0.3 billboard  0.3 allure )
```

1.5 Estimation

INQUERY, like most statistical systems, relies on a *tf.idf* formula for estimating the probability that a document is about a concept. Unlike many systems, INQUERY starts with a default probability and then adjusts it based on evidence. The formula used in recent TREC and TIPSTER experiments to determine the belief due to the occurrence of query term Q in a document was:

$$\mathrm{bel}_{\mathrm{term}}(Q) \;=\; d_b + (1 - d_b) \cdot \left(d_t \cdot H + (1 - d_t) \cdot \frac{\log(tf + 0.5)}{\log(max\_tf + 1.0)} \right) \cdot \frac{\log(\frac{C}{df})}{\log(C)} \qquad (1)$$

where

$$\begin{aligned}
tf &= \text{the frequency of term } t \text{ in the document,}\\
max_tf &= \text{the frequency of the most frequent term in the document,}\\
df &= \text{the number of documents in which term } t \text{ occurs,}\\
C &= \text{the number of documents in the collection,}\\
\mathrm{H} &= \begin{cases} 1.0 & \text{if } max_tf \leq 200 \\ \frac{200}{max_tf} & \text{otherwise} \end{cases},\\
d_t &= \text{minimum term frequency component when a term occurs in a document,}\\
d_b &= \text{minimum belief component when a term occurs in a document.}
\end{aligned}$$

In this formula, d_b is the belief in a document when there is no evidence ("default belief"). In recent TREC and TIPSTER experiments, d_b and d_t were set to 0.4.

The penalty H was introduced prior to TREC-1 to counter a slight bias towards long documents. This penalty was effective, if not pleasing theoretically. It prevented INQUERY from being biased unduly towards long documents, but still allowed them to be retrieved. (Only four of the TREC-2 systems retrieved more relevant Federal Register documents than did INQUERY [7].)

This year the document retrieval ("ad-hoc") experiments were conducted using a new estimation technique. The estimation experiments involved new forms of concept probability calculation. This probability is based on the *tf.idf* weights used in many systems, and these experiments incorporated factors such as document length. The penalty H was eliminated, the default belief (d_b) was based partly upon document length, and the default tf component (d_t) was based partly upon the frequency of the term. The adjustments to Equation 1 were:

$$H = 1.0 \tag{2}$$

$$d_t = 0.3 * \left(1.0 - \frac{df}{C}\right)^2 \tag{3}$$

$$d_b = \begin{cases} 0.4 - \frac{L}{1{,}000{,}000} & \text{if } (L < 400{,}000) \\ 0.0 & \text{otherwise} \end{cases} \tag{4}$$

where L = the length of the document (in words).
Our hypothesis was that the estimation technique should be more tolerant of not observing a query term in a short document than in a long document. Likewise, it should be more tolerant of not observing infrequent terms in a document.

These adjustments to the estimation formula were tested against more than just the TREC document collection. In experiments prior to TREC-3, they were found to yield small improvements at all levels of recall on the CACM (2 query sets), West FSupp (2 query sets), and NPL document collections.

2 Description of the Routing Experiments

The routing experiments incorporated two major changes. These were the inclusion of proximity-based features and the use of weights based on Rocchio [1]. Both phrase-level and paragraph-level proximities were considered. This means that significantly co-occurring words in both 5 word windows and 50 word windows were used as features in the automatically constructed queries. The Rocchio weights are based on averaging the *tf.idf* probabilities in both the relevant and non-relevant documents.

The INQ103 query set was created automatically from the original TIPSTER topics (100-150) and the relevance judgements from Volumes 1 and 2. The original queries were expanded using words and phrases from the relevant documents and reweighted.

The new features that were added to the queries were the 32 words with the highest weights in the relevant set, 10 proximities based on 5 word windows, and 20 proximities based on 50 word windows. Features were weighted using an approach based on the Rocchio formula developed for the vector space model [1]. The three primary differences between these weights and the standard Rocchio weights are:

1. The query is a weighted sum of three components,

```
#wsum(1.0 2.0 (original query) 4.0 (new word terms)
          1.0 (new proximity terms))
```

2. Since the top level of the query already contains the original query, the Rocchio computation only includes average weight in the relevant and non-relevant documents;

3. The term weights are computed using the weight functions INQUERY uses to estimate beliefs, and the the normalization uses document length statistics instead of the cosine normalization;

4. Although the entire relevant set is used, only the top-ranked n non-relevant documents are considered, where n is the number of relevant documents for the query, and the ranking is done using the original query.

The "manual" routing experiment (INQ104) was done by a weighted combination of the automatically produced queries with a manually produced set of queries generated for the ad-hoc experiments in the previous TREC experiments. No manual changes were made by examining relevant documents. The automatically produced queries were given twice the weight of the manual queries.

3 Description of the Spanish Experiments

The Spanish retrieval experiments focused on evaluating the effect of morphological processing. Only small modifications were required to apply INQUERY to a collection of Spanish documents. A stopword list was developed manually (1 week), document indexing had to be adjusted (1 week), a Spanish word stemmer had to be developed (5 weeks), and the graphical user interface had to be modified (6 weeks).

The word stemming algorithm is a Porter-stemmer in spirit [9]. Each stage of the Spanish algorithm is somewhat more complex than the English algorithm, because Spanish has more verb conjugation than does English, and the basic algorithm is augmented with a large table of exceptions, because Spanish has more irregular verbs than does English. Decisions about which verbs to handle with exceptions and which to handle with general rules were based upon Spanish verb data supplied by New Mexico State University.

The processing of Spanish topics to queries was accomplished by the same software used for English. Phrase recognition was disabled because we had no Spanish part-of-speech tagger. The stop-phrase heuristics were not translated to Spanish, so stop-phrases were not discarded.

4 Ad-Hoc Results and Discussion

Two sets of results, INQ101 and INQ102, were evaluated in the ad-hoc document retrieval evaluation.

The INQ101 results were based on completely automatic processing of the TREC topic statement into a query, automatic query expansion, use of passage-level and document-level evidence, and adjustments to INQUERY's estimation formula.

INQ102 was a semi-automatic experiment in which a user was allowed to edit the INQ101 query prior to running it. As in the past, the user was restricted to deleting query terms, grouping query terms with proximity operators, adjusting query term weights, and adding query terms from the Narrative. No other kind of modification was permitted.

The official results for INQ101 and INQ102 are summarized below.

Query Type	**Average Precision**			
	5 Docs	**30 Docs**	**100 Docs**	**11-Pt Avg**
INQ101	.64	.57	.44	.37
INQ102	.74 (+15.5%)	.63 (+11.0%)	.49 (+11.5%)	.42 (+15.5%)

Limited user modification of INQ101 produced a very significant 15.5% improvement in average precision. Most of this improvement appears to be due to 1) deleting useless query terms introduced by query processing or PhraseFinder, and 2) grouping query terms with proximity operators. Clearly there remains room for improvement in automatic query processing techniques.

After TREC-3, experiments were done to investigate the effect of the changes made this year. Our simple query processing improved precision very significantly at almost all levels of recall, as shown in the table below. The average precision obtained with query processing was 84.2% higher than the average precision obtained by using raw words alone.

Recall	Precision (50 queries)						
	Raw Words (RW)	RW W/O Stop & Not Phrases (SN)		SN With Field Weights (FW)		FW With #PHRASE op (PO)	
0	45.3	61.8	(+36.3)	78.8	(+74.0)	79.6	(+75.8)
10	28.2	40.1	(+41.9)	50.1	(+77.5)	57.0	(+102.0)
20	24.3	33.1	(+36.2)	40.3	(+66.2)	46.6	(+92.0)
30	21.6	29.7	(+37.3)	33.7	(+55.6)	40.1	(+85.5)
40	18.0	24.9	(+38.7)	28.0	(+55.6)	33.8	(+88.1)
50	15.7	21.3	(+35.7)	23.9	(+52.0)	27.7	(+76.5)
60	13.1	18.0	(+38.0)	20.3	(+55.8)	22.7	(+73.7)
70	10.6	14.7	(+37.8)	17.1	(+60.4)	18.4	(+72.8)
80	7.9	11.3	(+43.8)	13.1	(+66.9)	13.3	(+68.7)
90	4.5	7.2	(+60.3)	9.0	(+101.3)	9.4	(+110.4)
100	0.5	0.8	(+60.9)	0.9	(+99.8)	0.7	(+47.1)
avg	17.2	23.9	(+38.6)	28.7	(+66.3)	31.8	(+84.2)

Query expansion with PhraseFinder proved to be a good idea. As the table below shows, the inclusion of PhraseFinder terms in the query yielded a 9.6% improvement in average precision. Passage retrieval was also a good idea. Combining passage-level and document-level evidence produced a 15.7% improvement in average precision.

Combining PhraseFinder query expansion and passage retrieval led to the best performance of all, a 19.8% improvement over the baseline query processing. The improvement of the combination over passage retrieval alone was smaller than expected, suggesting that their are better ways of combining these two sources of evidence.

Recall	Precision (50 queries)						
	Query Processing (QP)	QP With PhraseFinder (PF)		QP With Passage (PS)		QP and PF and PS (All)	
0	79.6	79.4	(−0.2)	83.8	(+5.2)	80.6	(+1.2)
10	57.0	59.2	(+3.9)	61.7	(+8.1)	61.4	(+7.6)
20	46.6	50.3	(+8.0)	54.9	(+17.7)	54.1	(+16.1)
30	40.1	44.4	(+10.7)	47.2	(+17.6)	48.4	(+20.7)
40	33.8	37.4	(+10.7)	41.3	(+22.0)	43.0	(+27.1)
50	27.7	32.7	(+18.1)	34.0	(+22.8)	37.0	(+33.4)
60	22.7	27.6	(+21.6)	28.4	(+25.4)	31.6	(+39.4)
70	18.4	22.2	(+20.7)	23.2	(+26.5)	26.6	(+44.7)
80	13.3	16.8	(+26.7)	17.1	(+29.0)	20.6	(+55.5)
90	9.4	11.6	(+22.6)	11.6	(+22.7)	13.4	(+42.0)
100	0.7	1.2	(+73.5)	1.2	(+74.2)	1.9	(+177.0)
avg	31.8	34.8	(+9.6)	36.8	(+15.7)	38.1	(+19.8)

The change to the estimation formula, described above, had no significant effect on INQUERY's performance in the ad-hoc document retrieval evaluation. There was a +1.3% improvement in average precision over 11 recall points, but most of the improvement occurred at high (> 40%) recall.

5 Routing Results and Discussion

Two sets of results, INQ103 and INQ104, were evaluated in the document routing evaluation. The INQ103 results were based on completely automatic processing of the TREC topic statement and relevance judgements into a query. INQ104 was a combination of the INQ103 query and a query modified manually. The official evaluations are summarized below.

Query Type	**Average Precision**			
	5 Docs	**30 Docs**	**100 Docs**	**11-Pt Avg**
INQ103	.67	.58	.46	.38
INQ104	.62 (−1.9%)	.56 (−2.4%)	.45 (−3.0%)	.39 (+1.1%)

As in previous TREC experiments, the combination of automatic and manually-modified query sets was not helpful. This result is not surprising, given the improved weighting in INQ103 and the mediocre performance of the manually-modified query set in last year's TREC evaluation.

6 Spanish Results and Discussion

Two sets of results, SIN001 and SIN002, were evaluated in the ad-hoc Spanish document retrieval evaluation.

The SIN002 results were based on completely automatic processing of the TREC topic statement into a query. The processing performed was a subset of the processing applied to English topics. For example, there was no noun-phrase recognition, and no stop-phrase removal. The various fields of the topic statement were weighted as in English.

SIN001 was a semi-automatic experiment in which a user was allowed to edit the SIN002 query prior to running it. As with English, the user was restricted to deleting query terms, grouping query terms with proximity operators, adjusting query term weights, and adding query terms from the Narrative. No other kind of modification was permitted.

The official evaluations are summarized below.

Query Type	**Average Precision**			
	5 Docs	**30 Docs**	**100 Docs**	**11-Pt Avg**
SIN002	.74	.69	.60	.51
SIN001	.78 (+5.4%)	.67 (−2.7%)	.53 (−11.6%)	.42 (−17.5%)

Limited user modification of SIN002 produced improvements at very low recall, but a significant 17.5% loss in average precision. The relative inexperience of the user making the changes may have been a factor.

One concern was whether the Spanish word stemmer, which was developed in about five weeks, helped or hindered performance. The table below shows that it helped, providing a 9.4% improvement in average precision on the SIN002 query set.

Although the results for Spanish ad-hoc document retrieval appear reasonable, there is much room for improvement. A Spanish part-of-speech tagger would enable more sophisticated query processing, as would a Spanish stop-phrase recognizer. In principle, there is no reason why the complete English language query processing and query expansion techniques can't also be applied to Spanish. However, one practical near-term obstacle will likely be retrieval of large numbers of unjudged documents, due to the small number of systems participating in the Spanish evaluation this year.

Recall	Precision (25 queries)		
	SIN002 Unstemmed	SIN002 Stemmed	
0	86.5	86.9	(+0.4)
10	71.7	75.2	(+4.8)
20	65.3	69.6	(+6.7)
30	59.8	64.8	(+8.2)
40	54.5	60.4	(+10.9)
50	48.1	54.2	(+12.8)
60	42.7	49.1	(+15.0)
70	37.2	42.3	(+13.6)
80	29.2	33.0	(+13.1)
90	16.4	22.7	(+38.7)
100	3.5	5.1	(+44.4)
avg	46.8	51.2	(+9.4)

7 Summary

The themes guiding our work this year remain unchanged from previous years: We believe in highly structured queries, sophisticated query processing, and in combining multiple sources of evidence. The evaluation this year showed continued progress on each of these fronts.

INQUERY's heuristics for processing TIPSTER topics into queries continue to evolve and improve, providing a relatively high baseline from which to start considering other improvements. The ad-hoc document retrieval queries have become increasingly structured. Although no additional structure was introduced into the routing queries, their weights were improved by a switch to Rocchio weighting.

Two additional sources of evidence were introduced into the ad-hoc document retrieval queries this year: passage-level evidence, and query expansion using PhraseFinder. These provided significant improvements over the baseline query processing methods alone, and probably compensated to some extent for the loss of the "Concepts" field in the TREC topics.

The value of the Spanish evaluation was in showing how quickly INQUERY can be applied to new languages. Previous efforts have demonstrated that INQUERY can be used with Japanese [6] and Finnish, but they required significant effort. The Spanish effort required about seven weeks of work on the retrieval system, most of it to develop a Spanish word-stemmer.

Acknowledgements

We thank Lisa Ballesteros, Paul Booth, Stephen Harding, Tom Kalt, Zhihong Lu, and Jay Ponte for their assistance in the work described here. This research was partially supported by the NSF Center for Intelligent Information Retrieval at the University of Massachusetts, Amherst.

References

[1] C. Buckley, J. Allan, and G. Salton. Automatic routing and ad-hoc retrieval using SMART: TREC-2. In D. Harman, editor, *Proceedings of the Second Text REtrieval Conference (TREC-2)*. National Institute of Standards and Technology Special Publication 500-215, 1994.

[2] J. P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 302–310, Dublin, Ireland, 1994. Association for Computing Machinery.

[3] J. P. Callan and W. B. Croft. An evaluation of query processing strategies using the TIPSTER collection. In R. Korfhage, E. Rasmussen, and P. Willett, editors, *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 347–356, Pittsburgh, PA, June 1993. Association for Computing Machinery.

[4] J. P. Callan, W. B. Croft, and S. M. Harding. The INQUERY retrieval system. In *Proceedings of the Third International Conference on Database and Expert Systems Applications*, pages 78–83, Valencia, Spain, 1992. Springer-Verlag.

[5] W. B. Croft, J. Callan¡, and J. Broglio. TREC-2 routing and ad-hoc retrieval evaluation using the INQUERY system. In D. Harman, editor, *Proceedings of the Second Text REtrieval Conference (TREC-2)*. National Institute of Standards and Technology Special Publication 500-215, 1994.

[6] Hideo Fujii and W. B. Croft. A comparison of indexing techniques for Japanese text retrieval. In *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 237–246, 1993.

[7] D. Harman, editor. *The Second Text REtrieval Conference (TREC2)*. National Institute of Standards and Technology Special Publication 500-215, Gaithersburg, MD, 1994.

[8] Y. Jing and W. B. Croft. An association thesaurus for information retrieval. In *RIAO 4 Conference Proceedings*, New York, October 1994.

[9] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.

[10] Howard Turtle and W. Bruce Croft. Inference networks for document retrieval. In Jean-Luc Vidick, editor, *Proceedings of the 13^{th} International Conference on Research and Development in Information Retrieval*, pages 1–24. ACM, September 1990.

[11] Howard R. Turtle and W. Bruce Croft. Efficient probabilistic inference for text retrieval. In *RIAO 3 Conference Proceedings*, pages 644–661, Barcelona, Spain, April 1991.

[12] J. Xu, J. Broglio, and W. B. Croft. The design and implementation of a part of speech tagger for english. Technical Report IR-52, Center for Intelligent Information Retrieval, Department of Computer Science, University of Massachusetts, 1994.