Use of Query Concepts and Information Extraction to Improve Information Retrieval Effectiveness

David O. Holmes
NCR Corporation
Rockville, Maryland
david.holmes@washingtondc.ncr.com

M. Catherine McCabe
Advanced Analytic Tools
Washington, D.C.
cmccabe@gmu.edu

David A. Grossman
US Government
Washington, DC
dgrossm1@osf1.gmu.edu

Abdur Chowdhury
Illinois Institute of Technology
Chicago, IL
abdur@grouplogic.com

Ophir Frieder[1]
Illinois Institute of Technology
Chicago, IL
ophir@csam.iit.edu

Abstract:

In TREC-7, we participated in both the automatic and manual tracks for category A. For the automatic runs, we included a baseline run and an experimental run that filtered relevance feedback using proper nouns. The baseline run used the short query versions and term thresholding to focus on the most meaningful terms. The experimental run used the long queries (title, description and narrative) with relevance feedback that filtered for proper nouns. Information extraction tools were used to identify proper nouns. For manual runs, we used predefined concept lists with terms from the concept lists combined in different ways. The manual run focused on using phrases and proper nouns in the query. We continued to use the NCR/Teradata DBC-1012 Model 4 parallel database machine as the primary platform and added an implementation on Sybase IQ. We again used the relational database model implemented with unchanged SQL. In addition, we enhanced our system by implementing new stop word lists for terms and selecting phrases based on association scores. Our results, while not dramatic, indicate that further work in merging information extraction and information retrieval is warranted.

1. Introduction

Our work for TREC-7 is a continuation of the work started in TREC-3 when we implemented an information retrieval system as an application of a relational database management system (RDBMS). We used unchanged Structured Query Language (SQL) to implement vector-space relevance ranking (Grossman95, Grossman96).

[1] This work was supported in part by matching funds from the National Science Foundation under the National Young Investigator Program (contract IRI-9357785).

TREC-4 work demonstrated the relational implementation on category A data and introduced the concepts-list approach in the manual runs. In TREC-5, we implemented relevance feedback. TREC-5 also used the relational approach for the Spanish, Chinese and Confusion tracks. For TREC-6, we expanded our relevance feedback methodology to include the lnc-ltc term weights (Singhal96) as well as feedback term scaling. During TREC-6 we explored the assumption that certain infrequently occurring terms with high collection weights may actually be artificially inflating the query-to-document relevance ranking scores. We continued that work this year with expanded stop lists and term thresholding. In addition, this year we combined information extraction techniques with information retrieval through the use of a relevance feedback filter based on IE (Information Extraction).

Our manual runs have focused on the concept approach to structuring queries. In TREC-4, we assigned the query terms in up to three concept lists and used general world knowledge to expand the query to include other similar terms not found in the topic. In TREC-5, we continued to use the concept lists and experimented with the use of manually assigned weights to the query terms as well as using manual relevance feedback to identify additional terms. For TREC-6, we augmented our prior work with inexact term matching and an automatically generated thesaurus based on term-to-term co-occurrence. This year, our manual run took a somewhat more structured approach than in years past, with the hope of automating some techniques. In particular, our manual run focused on using phrases and proper nouns to improve precision and recall. Seventy percent of the manual search elements were phrases, which produced an average precision of 0.3333. At the 10-document retrieval level, our precision was 0.64.

2. Prior Work

2.1 Implementation of an Information Retrieval System Using the Relational Model

The implementation of an Information Retrieval (IR) system using the relational model hinges on the use of a relation (table) to model an inverted index which is central to traditional IR systems. The inverted index stores each unique term or phrase from the collection and a list of all the documents containing each entry. The inverted index can also include frequency, offset, or other desired information. In the relational approach, this index is flattened or normalized and stored in a table, as shown in Figure 2.1. Queries can be implemented using standard structure query language (SQL) to find and rank all documents containing the query terms. Full details of the implementation can be found in Grossman97 and Lundquist97. For example:

Collection:

D1: water freezes when cold
D2: the water is cold today

Inverted Index:

Cold → D1, D2
Freezes → D1
Today → D2
Water → D1, D2
When → D1

Relation:

Cold	D1
Cold	D2
Freezes	D1
Today	D2
Water	D1
Water	D2
When	D1

- *Figure2.1* -

One benefit to using the relational model for IR is the ability to exploit parallel processing via the DBMS. All commercial DBMS systems offer a parallel version. We implemented an IR system using Teradata's RDBMS on a 4 processor DBC/1012 parallel processing machine. The Teradata DBC/1012 Database Computer is a special purpose machine designed to run a relational database management system using standard SQL.

2.2 Relevance Feedback in the Relational Model

Building on prior work in automated relevance feedback, our TREC-6 submission implemented automated relevance feedback within the relational model, using unchanged SQL. Prior work has shown that the best results are achieved when not all of the terms from the top documents are used (Lundquist97, Buckley95). For instance, the top terms (based on normalized *idf* value) from the relevant documents are added to the original query terms and the query is rerun. Buckley added the most frequently occurring 50 single terms and 10 phrases from the top 20 documents. Lundquist showed that using the best 10-20 terms and phrases from the top 5 –20 documents with a scaling factor of .5 for the terms being added to the query (multiplying the weight by .5) performed best.

2.3 Information Extraction

Our TREC-7 implementation of relevance feedback experimented with a fusion of information extraction and information retrieval. Information Extraction is the process of identifying instances of entities within text.

Examples of entities include: organization, location, person, date reference, time reference, etc. The TIPSTER sponsored Message Understanding Conference is dedicated to improving the technologies for entity extraction. Our work focused on a subset of entities – proper nouns. Numerous techniques exist for tagging parts of speech using natural language processing. We used the commercial product, INSO, as the basis for identifying proper nouns. (Inso97).

2.4 Term-Term Association Techniques used for Stop List generation

The stop list used in the IIT (Illinois Institute of Technology) system was expanded for TREC-7 based on the positive impact of a larger stop list used during experimentation for TREC-6. A manually generated stop word list for term processing which consists of approximately 3,500 words included based on high document frequency and variants (prefixes and suffixes, as we do not use stemming) of those terms with high document frequency. A second stop list, a subset of the term stop list, is used for phrase preprocessing. The idea is to index high quality phrases rather than merely locate adjacent word pairs. We eliminate from the phrase stop list those terms which have a high affinity with other terms in the collection that are found adjacent to the potential stop word. High frequency terms remain on the term stop list but because they qualify as existing in a high-association pair, they are removed from the phrase stop list and phrases containing that word are indexed. For instance, "big apple" might be retained while 'big' alone might be removed.

Three measurements determined the contents of the stop list. To measure the affinity of the first word in a phrase to the entire phrase, we computed:

$$a1 = \frac{b}{c}$$

where

a is the association measure
b is the term frequency of a phrase
c is the term frequency of the first word in a phrase
d is the term frequency of the second word in a phrase

To measure the affinity of the second part of the phrase to the phrase, we computed:

$$a2 = \frac{b}{d}$$

The overall association measure was computed as follows:

$$a3 = \frac{b}{((c+d)/2)}$$

Terms were removed from the stop list if *a3* greater than 0.0005 (we calibrated several different values and this was best) and either *a1* or *a2* was also greater than 0.0005 for one or more phrases in the collection. Figure 2.4-1 provides examples; italicized entries do not qualify as phrases. Given the word *higgledy,* there is a 0.94 probability it is followed by *piggledy* in the TREC6 collection, which indicates, by our metric, it is a content-bearing phrase. While *brothers owned* is not a phrase, *Ringling Brothers* is. While not empirically measured, a casual inspection of phrases with high association scores revealed a substantial percentage of proper nouns.

First Word	Second Word	Phrase TF	First TF	Second TF	a3
Higgledy	Piggledy	16	17	17	0.94000
Humpty	Dumpty	38	51	39	0.84000
Sherlock	Holmes	105	236	1,641	0.11000
World	Bank	7,813	176,487	172,982	0.05000
Orange	Juice	536	32,578	2,477	0.03000
Nuclear	Waste	1,745	83,479	34,811	0.03000
Big	Apple	106	73,665	4,984	0.00300
Ringling	Brothers	12	54	13,059	0.00200
David	Holmes	13	34,375	1,641	0.00070
Nuclear	Family	43	83,479	73,511	0.00060
Women	***Buy***	***13***	***59,844***	***43,870***	***0.00025***
Brothers	***Owned***	***3***	***13,059***	***43,939***	***0.00007***

Figure 2.4-1 Self-Association Examples

We compared several possible term and phrase stop lists using TREC-6 topics to determine our choice for TREC-7. The results are summarized in Figure 2.4-2.

Stop List	Average Precision	Relevant Retrieved
SMART	0.1694	2102
Fox	0.1744	2086
TREC-6 (GMU)	0.1816	2085
TREC-7 (new lists)	0.1905	2122

Figure 2.4-2 Comparison of Stop Lists on Average Precision

3. Implementation Details

3.1 Automatic Runs

3.1.1 Automatic Run 1 — Proper Nouns Filter Relevance Feedback

The IIT automated long run experimented with using proper nouns for relevance feedback only. The initial queries consisted of the title, description and narrative portions of the TREC-7 queries. Relevance feedback was used to augment the queries with the top proper nouns from the top documents.

In order to limit the number of terms added in relevance feedback, we developed a filter which limits the terms added in relevance feedback to the top 10 proper nouns, ranked by N**idf* where N is the number of times the term occurs in the top 20 documents. We used the INSO product to identify the noun phrases (including single-term nouns) from a subset of the TREC7 collection. A crude assumption was made that any capitalized noun phrase was a proper noun. This list was then used as a filter to select terms for relevance feedback. First, the unaltered long queries were run. The top 10 documents were retrieved, and the top ten index entries for those documents were identified (based on N*idf). That list of candidates was then compared against the extracted proper nouns. If found on the list, the term was added to the original query. In this way, ten or fewer terms were added to each query. Because the IIT system only indexes single terms and term pairs, only those noun phrases were actually found in the index. INSO identified 367,142 noun phrases from the training subset of theTREC-7 corpus. The benchmark against TREC-6 is shown in figure 3.1-1.

Test	Average Precision	Relevant Retrieved
No feedback	0.1905	2122
INSO phrase feedback	0.1925	2164

Figure 3.1-1 IIT Automatic Run 1 TREC6 Benchmark

3.1.2 Automatic Run 2 — Baseline

The second automatic run formed our baseline, using the title portion of the TREC-7 queries and no relevance feedback. This run was implemented using Sybase IQ and took advantage of the bit-wise indexing available in that database. We implemented query thresholding, which eliminated low *idf* terms from the query unless it was the last remaining term in the query (Grossman97). In addition, we thresholded the index used for this run and eliminated the most common terms and phrases. This resulted in a very fast system – because the most frequent terms were removed. In addition, the index size was reduced by half. The results were not good, however, indicating that we may have gone too far in the cut-off point used for thresholding.

3.2 Automated Concept Experiment

This experiment was conducted after submission of our official runs for TREC-7. Automatic tests against TREC-6 and manual runs against TREC-7 indicated improved precision and recall with the implementation of concepts which limit the size of the answer set for each topic. To do this automatically, we required each retrieved document to contain at least one term or phrase from the Title concept. Terms and phrases from the Description and Narrative sections contributed to the relevancy score, but did not identify new documents.

The IIT system does not use stemming and the fallout from using terms and phrases from only the Title section for document qualification is dramatic. To minimize the problem, the Title concept was automatically expanded, based on several criteria. The scoring only concept was also automatically expanded to improve ranking. Figure 3.2-1 summarizes the results of the experiment with TREC-6 topics.

TREC-6 Test	Average Precision	Relevant Retrieved
Long	0.1905	2122
Title Concept, Long Scoring	0.2559	2273

Figure 3.2-1 Automated Concepts Experiments on TREC-6 data

3.3 Manual Run

3.3.1 Manual Run Implementation Details

We conducted failure analysis of our TREC-6 results and found that when the statistical approach to relevance feedback added phrases and proper nouns, precision and recall generally improved (Lundquist98). In particular, topic 311 performed well and contained names from high profile industrial espionage cases. To produce the manual queries, we followed a two-step process. First, the analyst read each topic and spent approximately one to ten minutes building an initial query. After completing the initial queries, the analyst briefly scanned the results and examined a few documents in detail for each topic. The analyst read high-ranking documents and others with intriguing headlines, adding promising terms and phrases to the queries. In some cases, the analyst issued a query to find frequently occurring phrases within a selection of five to ten potentially interesting documents. The entire process took approximately 15 to 30 minutes per topic.

Our manual submission concentrated on using phrases and proper nouns to improve precision and recall. Queries consisted of 908 phrases and 389 single terms. Of the total, 541 were proper nouns including 235 names of people.

Similar to last year, we implemented one or more mandatory concepts for each topic. Queries selected documents only if the documents contained one or more entries from each mandatory query concept. For a few topics, we attempted to eliminate non-relevant documents by adding ‘negative concepts’ — terms which must NOT be in the document for the document to be selected. However, our implementation of this feature contained an error that resulted in these terms actually being *required*. For topic 351, documents should have been dropped if "fishing boats" occurred within the body of text, instead those documents were actually added to the result set. Finally, to enhance ranking without qualifying additional documents, we implemented a *scoring-only* concept. If a document contained a term from this concept, its ranking would be adjusted.

3.3.2 Manual Run Failure Analysis

Our official manual run was above the median on 30 topics and below on 17. Figure 3.3-1 shows how the manual results differ from the median.

The SQL used in the manual run had an error that applied the negative concept terms to all topics, instead of limiting them to their assigned topics. This error degraded the performance of 28 topics. Interestingly, the error actually improved 21 topics and on topic 363 the mistake pushed the total recall to 100% — the fix lost one of the relevant documents by ranking it lower than the top 1000. Overall, average precision degraded by 5% because of the system error regarding negative concepts.

Topic 389 demonstrates the potential value of using proper nouns and phrases. While the median was 0.0305, the IIT system averaged 0.4985. Proper nouns such as South Africa, Carlos Cardoen, Christopher Drogoul, Edward Bush, Jonathan Pollard contributed to the success of this query.

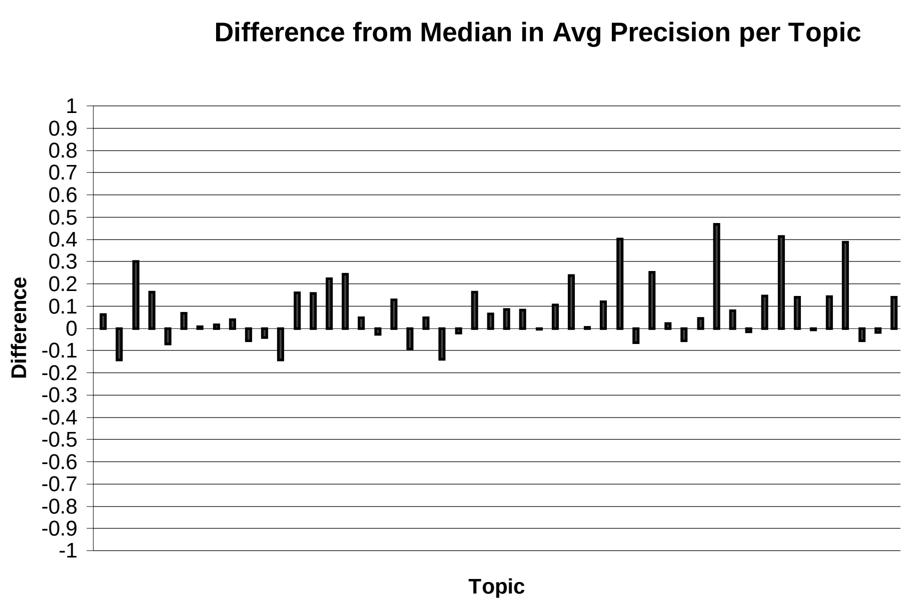

Figure 3.3-1: IIT Manual Run Difference from the median

4. Results

Figure 4-1 summarizes the results of our TREC-7 tests. An error in the SQL used for the manual run lowered our official results. The corrected SQL, using the exact same search terms, was run again and the results are included.

	iit98au1 (Long)	iit98au2 (Short)	Iit98ma1 (Manual)	iit98ma1 (Fixed)	Auto Concept
Avg. Precision	0.1929	0.1459	0.3333	0.3511	0.2091
Precision at 10 Documents	0.4500	0.3200	0.6400	0.6540	0.4740
Relevant Retrieved	2505	1676	2793	2914	2495
Above Median (Avg. Precision)	24	10	30	34	27
Below Median (Avg. Precision)	26	40	17	16	23

Figure 4-1: IIT Trec7 Results Summary

5. Conclusions and Future Work

For TREC-7, we focused on integrating Information Extraction and Information Retrieval through the use of a relevance feedback filter. In addition, we refined our use of concepts in the manual runs. Our results show promise in the use of phrases and proper nouns. Using term-term association scores to control the preprocessing of phrases improved our system over last year. This year, we took our initial steps in implementing information extraction tools to automatically expand queries to include proper nouns. The marginal improvement experienced this year is probably indicative of the preliminary implementation using only proper nouns from a sample of the corpus.

Our future challenges include automating the techniques used in our manual process. Our initial tests indicate that some degree of automatic concept detection is possible in the TREC environment. Post-TREC-7 experimentation with automatically selected terms and phrases for a mandatory concept and a scoring concept resulted in improving average precision by 8.9%. Further use of information extraction and term-term associations are candidates for further concept refinement. In addition, further integration of information extraction in relevance feedback is needed to move beyond proper nouns and experiment with the use of entities as feedback filters.

6 Acknowledgments

We would like to thank James Dyer and Bret Bailey of Sybase for their support in the IQ implementation.

References:

(Buckley95) Buckley, C. A. Singhal, M. Mitra, and G. Salton, “New Retrieval Approaches Using SMART: TREC-4,” sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1995.

(Fox90) Fox, Christopher. A Stop List for General Text. SIGIR Forum, (v. 24, no. 1-2) 1990, p. 19-35.

(Grossman95) Grossman, D., D. Holmes, O. Frieder, M. Nguyen, and C. Kingsbury, “Improving Accuracy and Run-Time Performance for TREC-4,” Proceedings of the Fourth Text REtrieval Conference (TREC), sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1995.

(Grossman96) Grossman, D., C. Lundquist, J. Reichert, D. Holmes, and O. Frieder, “Using Relevance Feedback within the Relational Model for TREC-5,” Proceedings of the Fifth Text REtrieval Conference (TREC), sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1996.

(Grossmam97) Grossman, D., D. Holmes, O. Frieder, and D. Roberts, “ Integrating Structured Data and Text: A Relational Approach,” *Journal of the American Society of Information Science*, January 1997.

(Inso97) Inso Intelligent Parts of Speech Tagger. Inso Corporation. 1997.

(Lundquist97) Lundquist, C., D. Grossman, O. Frieder, and D. Holmes, “A Parallel Implementation of Relevance Feedback using the Relational Model,” *Proceedings of the World Multiconference on Systemics, Cybernetics, and Informatics*, July 1997.

(Lundquist98) Lundquist, C., D. Holmes D. Grossman O. Frieder. Expanding relevance feedback in the relational model. NIST Special Publication 500-240, pages 489-502, August 1998.

(Singhal96) Singhal, A., C. Buckley, and M. Mitra, “Pivoted Document Length Normalization,” *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Ed. Hans-Peter Frei, Donna Harman, Peter Schauble and Ross Wilkinson, SIGIR Forum, August 18-22, 1996.