READWARE® TEXT ANALYSIS AND RETRIEVAL IN TREC 7

By Tom Adi, O. K. Ewell and Patricia Adi
Management Information Technologies, Inc.
Email: mitioke@readware.com

ABSTRACT

This paper describes Management Information Technologies, Inc.'s (MITi) first involvement in the TREC program. We limited our participation to manual adhoc although our multilingual algorithms can be used for automatic query generation and refinement and are suited for most TREC tasks.

We used our commercially available text analysis and retrieval Readware technology to perform the manual adhoc task of finding the documents relevant to fifty specified topics in a pool of more than half a million documents. Readware uses concepts (groups of related words), superconcepts (groups of concepts), Readware query elements (query building blocks) and document subjects to form queries. This is complemented by word search, phrase search and Boolean logic.

One MITi analyst performed the task. She formulated an average of 18 queries per topic. The queries were derived intuitively from topic specifications (title, description and narrative). First, a baseline pool of documents was identified for every topic using a few simple queries. Then the analyst queried the baselines using as often as possible Readware query elements related to elements of the topic specification. On average, few hits were returned per query. The analyst also had the advantage of seeing the exact responsive text spots highlighted in every hit document. Queries were adjusted and expanded using information from the neighborhood of the highlighted hit spots. There was no intrinsic ranking of hits. All hits were full semantic matches. The hits were ranked higher after the fact if the queries contained more items.

In the "Best Manual Adhoc" figure of the TREC 7 evaluation results, MITi's graph is above all other participants' graphs at most points.

1. INTRODUCTION

MITi has been developing text analysis and retrieval techniques under the trade name Readware for over a decade. This is our first participation in TREC. We submitted one run in the manual ad hoc category.

Readware currently uses the following conceptual sets to analyze text:

a) a few thousand concepts (groups of related words)
b) a few dozen superconcepts (groups of concepts)
c) a few hundred query elements (query building blocks) composed of superconcepts, concepts, words and phrases
d) 46 document subjects identified by concepts

Readware query elements include:

a) query helpers: frequent questions such as "who," "where," "why," "what does it mean"
b) Readware topics: useful extended concepts such as courts, leaders, safety, medicine, military and business
c) issues people use to make critical decisions, such as "emerging needs," "potential trouble," and "checking on those in charge."

There are three basic search strategies: word search, concept search and superconcept search. Readware's selectable query elements simplify the art of asking questions.

Users may mix strategies using a different strategy for every item. A variable-size sliding search window scans each document for certain words, phrases, concepts and superconcepts. The window size (context size) can be set in the query to values ranging from one tenth the query size to 20 times the query size.

Queries may also contain document-level inclusion or exclusion of words and phrases. And finally, Readware incorporates Boolean logic to combine words, phrases, concepts and query elements into a compound query.

MITi participated in the manual adhoc task. One MITi analyst used Readware to prepare and query over 500,000 test documents. The goal was to identify all the documents discussing each of the fifty TREC 7 topics. Hit documents were required to be ranked and a maximum of 1000 hit documents were expected per topic.

Since all Readware hits must have a complete set of full semantic relations with the query and Readware no longer ranks hits by semantic points, we only looked for "good hits" without ranking. After establishing a rather small baseline set of documents for every topic, queries made to the baseline returned few hits on average. The exact hit spots were highlighted and the analyst was able to judge the hits rather quickly. Queries were refined and in the end, most of the hit documents we delivered were already judged likely relevant from the analyst's point of view.

We delivered for evaluation a total of 5898 ranked hit documents. For the purposes of TREC evaluation, we ranked the hits by the complexity of the queries used. The more items and positions the generating queries contained, the higher the hit rank. We did not rank hit documents higher if they contained more than one hit spot.

2. DATA PREPARATION

We used a Pentium II (266 MHz) with 128 MB of RAM and a 4 Gigabyte disk. A fully automatic data preparation took about 8 hours of CPU time.

Once the TREC 7 files were decompressed, they were ready for automatic processing by Readware. It took some minor programming to exclude certain fields such as subject, headline and header. Our compiler split the files into documents using the <DOC> and <DOCNO> tags. This was done without physical duplication by keeping track of document lengths and their positions in the original files. Our default tag filter made sure that tag contents were skipped.

Every document was analyzed to determine the positions of words, concepts, phrases and query elements. Identifying Readware query elements meant asking about a million questions to every document using a variable-size sliding search window. Document subjects were identified using concept frequencies at the tops of documents.

The results of the analysis were stored in 4 files:

docs._ (42 MB):	table of vital data per document: document number, file number, position in file, document subject, document issues, etc.
list._ (71 k):	TREC file list containing documents
sigs._ (931 MB):	Readware signature database of positions of words, concepts and query elements in every document.
optdx._ (155 MB):	Optimized index

The text analysis consulted the 2 MB Readware Concept Base which consists of several files.

3. QUERY CONSTRUCTION

To save search time, we first limited the scope of search to a small pool of topic-related documents. To identify such documents, we asked simple questions.

> For TREC topic 372 "Native American casino," we searched for the words casino or gambling combined with Indians, Native Americans, tribes or reservations. The search identified about 260 TREC documents. This became the pool for further searches and we called it a "baseline."

Most TREC queries were asked within a baseline pool, and sometimes the baseline was expanded during the search. Setting a baseline also established the maximum possible number of relevant documents. Baselines were very efficient. For example, searching the baseline of topic 372 for the word "casino" brought a majority of good hits.

Then, we looked at TREC topic specifications composed of title, description and narrative and tried to identify the basic elements.

> TREC topic 372 "Native American casino" contains the elements: growth of gambling, social implications, economic effects on community and tribe, and legal aspects related to tribal autonomy.

Our search got off to a quick start by clicking on the Readware query elements which corresponded to the basic elements of the TREC topic. Experience showed this was the fastest way to find relevant documents.

> The query "Indian casino" combined with query helper CONSEQUENCES or query helper WHY found implications and effects. Combined with the Readware topic POLICE or the collective issue ALL ISSUES, this search captured many hits related to community disruption. We did not have to ask specific questions about social implications, economic effects and legal aspects.

Refining queries by combining (AND) increased precision and recall.

> The queries "Indian gambling" and "tribal casino" within the baseline pool brought a mix of good and bad documents. But "Indian gambling" AND the Readware topic POLICE brought 7 clean hits. And "tribal casino" AND the query helper WHY found 8 good hits.

Queries were also refined by excluding (NOT) unwanted words, concepts, phrases, query elements (also at document level) and excluding document subjects.

> "Indian gambling affairs" but NOT the word (Trump) brings 14 clean results.

Using alternative query formulations and strategies made the search more exhaustive.

> Alternative queries included "reservation gambling," "Indian gambling" and "Native American casino." The queries can be searched with word search or concept search. The search "window" (context size) may be small or large. Mixed strategies are possible in one query-- partly word, partly concept and partly document wide search.

Compound queries show relatedness and focus the search process.

> Casino/gamble/gaming AND (query helper PROBLEMS/FAILURE OR query helper SUCCESSES)

In the neighborhood of the exact responsive text spots highlighted by Readware, we found words, concepts and elements which we used in more queries.

> We learned from the text around highlighted responsive spots that Native Americans from the Mohawk tribe ran a casino. When we asked the simple question "Mohawk" within our baseline pool, we got 4 clean hits.

A series of focused queries which bring back a manageable number of results (say, from 1 to 50 at a time) are more satisfying for an analyst to work with than a strategy that requires him to sift through hundreds or thousands of irrelevant responses for a few good ones.

> A total of 14 questions was asked for TREC topic 372. The average number of hits per query was 6. The maximum number of hits was 17; the minimum was 1. Good document hits were marked. The analyst stopped asking questions when no more relevant new documents were discovered.
>
> Out of the 46 hits we delivered for this topic, 43 were judged relevant. Judges found a total of 49 relevant documents.

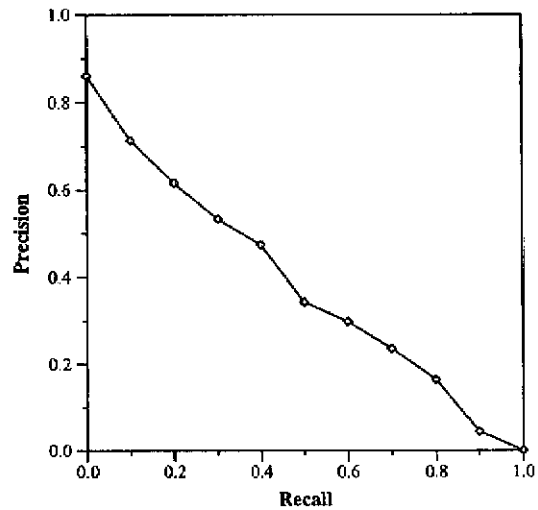
For all 50 topics, we constructed a total of 918 queries, an average of 18 queries per topic. We used document subjects and Readware query elements over 650 times in the queries, i.e. in 7 out of 10 queries.

4. PERFORMANCE

In the TREC 7 evaluation results, MITi is listed as one of the best manual ad hoc runs. Our graph (t7miti1) appears above all other graphs at most points of the comparative recall/precision figure.

MITi scored the highest R-Precision (precision after R documents retrieved) at 0.4392 (second-highest is Claritech at 0.4140). We achieved the second-highest average precision over all relevant documents at 0.3675 (just below Claritech's 0.3702).

MITi delivered the smallest number of retrieved documents, 5,898 (followed by University of Waterloo's 16,617) but we had the second-highest number of unique contributions, more than 160 (following Waterloo's 200 or so).



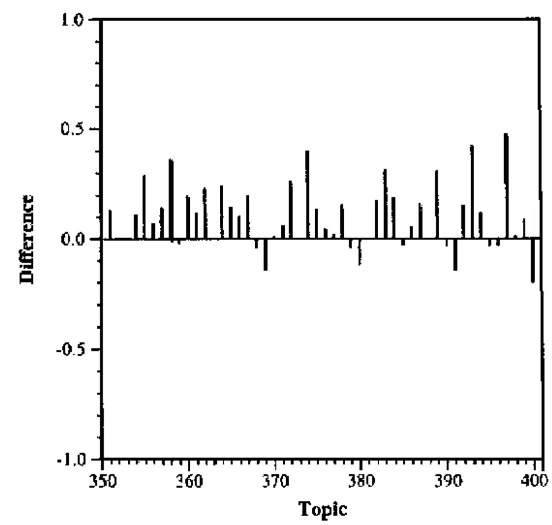
Recall-Precision Curve

HIGHLIGHTS

Summary Statistics	
Run Number	t7miti1
Run Description	Manual
Number of Topics	50
Total number of documents over all topics	
Retrieved:	**5898**
Relevant:	4674
Rel-ret:	2520

Recall Level Precision Average and Selected Document Level Averages

Average Precision over all relevant documents	
Non-interpolated	**0.3675**
R-Precision (after Relevant docs retrieved)	
At 5 docs	0.6640
At 10 docs	0.6400
At 15 docs	0.6213
At 20 docs	0.5780
At 30 docs	0.5433
At 100 docs	0.3512
Exact	**0.4392**



Difference from Median in Average Precision per Topic