High Selectivity and Accuracy with READWARE's Automated System of Knowledge Organization

By Tom Adi, O. K. Ewell and Patricia Adi
Management Information Technologies, Inc. (MITi) [1]
Email: mitioke@readware.com
October 27, 1999

Abstract

READWARE performs a fully automatic text analysis that implements a system of knowledge organization based on *knowledge types*. A knowledge type is a set of instructions that identifies a set of *knowledge elements* in any text. Knowledge types include *concepts* (word sets), *topics* (an expandable hierarchical scheme of common knowledge types spanning politics, business, health, and so on), *probes* (investigative knowledge types), *issues* (knowledge types used in decisionmaking) and *document subjects* (traditional classification of documents by themes). An MITi analyst used this system to translate TREC topic specifications into highly selective queries (few hits per query) in two adhoc runs with high relevance rates (2019 / 3060 hits in the READWARE run and 2774 / 5785 hits in the READWARE2 run).

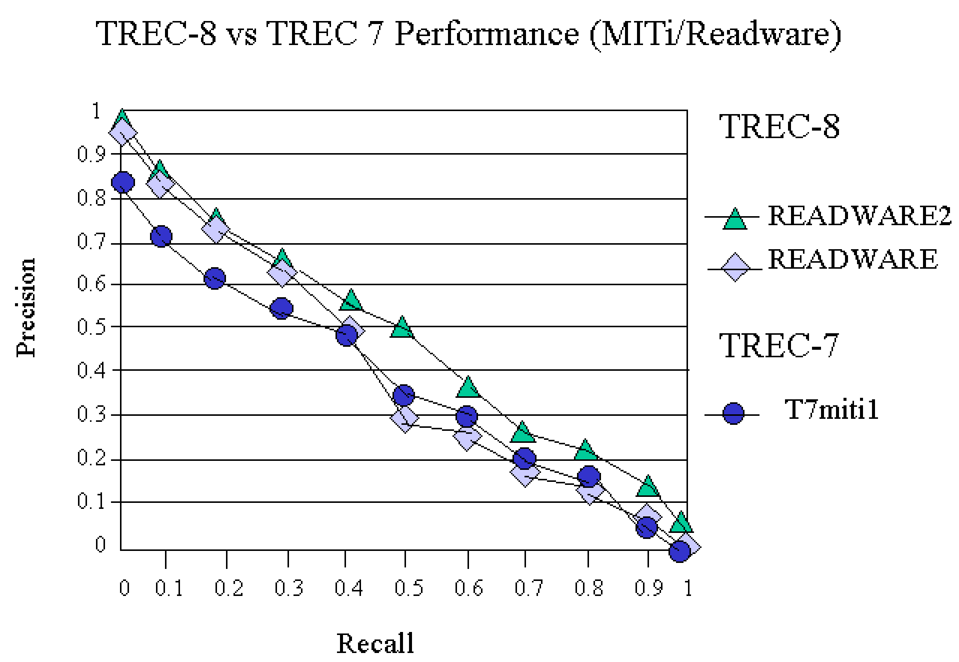

Figure 1.

[1] **Management Information Technologies, Inc.** has been developing software for automatic text analysis and search based on a system of knowledge organization since 1985. MITi's technology is marketed under the trade name **READWARE®**. The product line includes **ConSearch** for Windows workstations, the **IpServer** for internet/intranet applications and the Readware Software Development Kit for custom solutions.

1. Introduction

MITi is participating in the TREC for the second time with its READWARE technology. We used our product **ConSearch** to perform two manual adhoc runs. The adhoc task consists of finding the documents relevant to fifty specified topics in a pool of more than half a million documents.

READWARE is a text analysis technology based on a system of knowledge organization consisting of **knowledge types**. A knowledge type is a set of instructions (usually a set of queries) that identifies a certain set of **knowledge elements** in any text. **Concepts** are a network of basic knowledge types. A concept is a set of words that are seen as strongly related knowledge elements. A **superconcept** is a compound knowledge type consisting of several closely related concepts. Concepts and superconcepts are inspired by terms from ancient languages. READWARE offers the following advanced knowledge types:

1) **Document Subjects.** Document subjects reflect the traditional classification of documents by themes. Analysts can create their own **user subjects**.

2) **Probes**. Probes are investigative knowledge types such as *who/where, why, how often, success, growth* and *roots.*

3) **Issues**. Issues are knowledge types used in decisionmaking such as *trends, emerging needs, potential trouble, checking on those in charge* and *clash of interests.*

4) **Topics.** Topics form an expandable hierarchical scheme of common knowledge types. Analysts can easily construct their own **user topics**. The current scheme consists of the following topic areas that contain a total of 336 topics:

Current READWARE Topic Areas

- Way of Life
  - World View
  - Laws & Lawmakers
  - Courts
  - Those in Charge
- Communities & Relations
  - Family
  - Countries & Regions
  - International
- Basic Needs
  - Safety and Security
  - Health
  - Knowledge & Technology
  - Sources of Information (Media)
  - Environment
  - Housing
  - Transportation
  - Food
  - Energy
  - Business
- Culture & Recreation
  - Culture
  - Recreational Activities
  - Travel
  - Sports

READWARE also offers Boolean logic, sequence enforcement, context sizing, word search, phrase search and document-level search.

This year, we refined and extended our knowledge types, especially topics and probes. The number of topics doubled since last year. We also included many British spellings and terms.

There are three basic search strategies: word search, concept search and superconcept search. The document pool may be limited by specifying which document subjects are desired or undesired. Users may mix strategies using a different strategy for every query item. A variable-size sliding search window scans each document for certain words, phrases, concepts, superconcepts, topics, issues and probes. 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.

READWARE hits must have a complete set of semantic relations with the query. We ranked the hit documents by the complexity of the queries used. The more items and positions the generating queries contained, the higher the hit rank. But unlike last year, a hit document was ranked higher if there were multiple hit spots or a hit spot was in the headline or near the top of the document. READWARE highlights the exact hit spots in the documents.

2. Data Preparation

Knowledge type implementations are stored in the **ConceptBase** (5 MB for English). The READWARE text analysis module automatically locates all knowledge elements in the texts before the analyst sits down to make queries.

We used a Pentium III (450 MHz) with 256 MB of RAM and a 12 Gigabyte disk. A fully automatic data preparation (text analysis) took about 13 hours of CPU time.

TREC 8 files were decompressed and their end-of-line sequences were optimized using a utility program. The READWARE text analysis module then split the files in memory 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 tags were not analyzed. The text parts between the <subject> and </subject> tags were also not analyzed.

The READWARE analyst module scanned every document to determine the positions of names (non-concept words), concepts, topics, issues and probes and to identify document subjects. Locating the knowledge elements belonging to all READWARE knowledge types meant asking over a million queries to every document using a variable-size sliding analysis window.

Analysis results were stored in 3 files:

docs._ (68.5 MB): vital document info (document file, subject, issues and topics)

sigs._ (1.04 GB): signature database (positions of names and knowledge element in all documents)

optdx._ (176 MB): optimized index

3. Query Construction

One MITi analyst used READWARE's knowledge types (a few thousand concepts, a 336-topic hierarchy, 27 probes, 21 issues and 54 document subjects) and the advanced search features (Boolean logic, sequence enforcement, context sizing, word search, phrase search and document-level search) to refine and perfect the TREC topic specifications (title, description and narrative) and turn them into a more consistent and complete set of automatically executable queries.

READWARE indicates to the analyst what concept a certain word in the TREC topic specification belongs to so that she can search for the full concept rather than a single word if she chooses to do so. And whenever the specifications are ambiguous or incomplete, the analyst can navigate the knowledge type schemes and follow her hunches. She can expand her thoughts and fill in the gaps by testing queries made with indicated concepts, topics from the same topic area, similar probes, related issues and other document subjects. Extensive word expansion and thought expansion are already implicit in the knowledge type implementations.

The MITi analyst first constructed 65 *user topics*. These are READWARE topics that should be found in every document dealing with the TREC topic. Used as queries, they identify a baseline pool of documents for every TREC topic.

For example, here are the TREC specification and the READWARE user topics for TREC topic 420:

TREC Specification for Topic 420

Title: Carbon monoxide poisoning
Description: How widespread is carbon monoxide poisoning on a global scale?
Narrative: Relevant documents will contain data on what carbon monoxide poisoning is, symptoms, causes, and/or prevention. Advertisements for carbon monoxide protection products or services are not relevant. Discussions of auto emissions and air pollution are not relevant even though they can contain carbon monoxide.

User Topic 367: Carbon Monoxide

```
//The following line is the topic title
=aa carbon monoxide (367)
//Look for the words “carbon” and “monoxide”
//The phrase  “emission” (from narrative) not in the context
//Topic 106 (air pollution, from narrative) not in the context
!"emission"
carbon monoxide !T:106
```

User Topic 373: CO

```
//The following line is the topic title
=aa CO (373)
//Look for the phrase “CO” (case-sensitive)
//The words emission (from narrative) dioxide and founder
//should not be in the context
/""CO"
/+ !emission !dioxide !founder
```

The lines starting with “//” in the user topic boxes are comment lines that explain the queries and their relationship to the TREC topic specification. The phrase *air pollution* in the description is a READWARE topic title (topic 106).

In addition to the queries in the user topics, the analyst formulated 771 queries, an average of 14 queries per TREC topic (compared to 18 last year). She made queries by combining baselines or user topics with knowledge types and phrases related to the TREC topic specifications. Topic 445 (women clergy) needed the least number of queries, just two. Topic 401 (foreign minorities, Germany) required the most number of queries, 65.

To satisfy different styles of judging, the analyst made two runs. In the first run (labeled READWARE, *the stickler run*), she tried to be literal, making precise queries that included all the elements and satisfied all the conditions required in the topic specifications. In the second run (READWARE2, *the comprehensive run*), she incorporated the hits of the first run and added more hits using less precise queries. Here are two query examples for topic 420 (carbon monoxide poisoning):

Query 1

```
b: G:W P:5.0 poisoning H:2 S:1 S:2 !{"strychnine"
```

Query 2

```
G:C P:2.0 !{"air quality" !{chemical !{emission  S:37 T:367
```

Query 1 is meant for the stickler run (labeled READWARE). We are looking for the word *poisoning* (G:W means word search) in the context of size 5 (P:5.0) of the probe H:2 (what are the numbers, implements the description phrase “*how widespread*”) in documents from the baseline pool (b:) classified as having the subjects *Accidents/Crisis* (S:1) or *Crime/Police* (S:2) but not including the word *strychnine* anywhere. Document subjects were specified to fit the *poisoning* theme and to avoid getting ads or environmental documents that are excluded by the narrative. The query returned 12 hit documents that were all accepted as relevant by the judges.

Query 2 is meant as an addition to the comprehensive run (labeled READWARE2) that includes all the hits from the stickler run. We are looking for the user topic *Carbon Monoxide* (T:367) using concept search (G:C) in a context of size 2 (P:2.0) in documents from the complete pool which are classified as having the subject *Medicine* (S:37, suggested by the word *symptoms* in the narrative) but not including the phrase *"air quality"* or the words *chemical* or *emission* anywhere in the document. The query returned 7 hit documents none of which were accepted as relevant by the judges (the analyst was right to exclude them from the stickler run).

A total of 12 queries were made for TREC topic 420. We retrieved only 38 hit documents for this topic in the stickler run. 27 were judged relevant. The total number of relevant documents found by the judges in 13 runs for topic 420 was 33. In the comprehensive run, we delivered 59 hit documents to the judges (including 38 from the first run) out of which only 28 hits were judged relevant. This consisted of 27 hits from the first run and only one hit from the comprehensive run.

4. Performance

The stickler run (labeled READWARE) showed very high selectivity and retrieved a total of only 3060 hit documents. Of these hits, 2019 were judged relevant. The stickler run had an average relevance rate of 66%. The average precision (non-interpolated) of the stickler run is 40% (3% better than last year). The R-precision (exact) is about 45% (1% better than last year).

The comprehensive run (labeled READWARE2) had a relevance rate of 48%. We retrieved only 5785 documents out of which 2774 were judged relevant. The average precision (non-interpolated) of this run is high at about 47% (10% better than last year). The R-precision (exact) is high at about 51% (7% better than last year). In this run (based on an evaluation over 13 runs), MITi scored best average precision in 23 topics compared to only 8 topics last year.

Figure 1 on the first page shows MITi performance this year in both runs as compared to our TREC-7 run of last year.

Figure 2 below graphs READWARE's precision over X documents in both TREC 8 runs. This figure shows high precision in the first 30 documents retrieved in both runs.

5. Conclusion and Outlook

Knowledge managers and analysts both can enjoy direct access to highly relevant sets of documents retrieved by READWARE as predicted by the statistics in Figure 2 below and elsewhere. They can have high confidence that the text analysis performed by READWARE yields a larger pertinent measure of the "relevant whole" and that this reliably reduces the amount of information that needs to be read in further analyses. Any analyst would rather examine 3060 documents than 50,000 if given the choice, when they can be assured that the lower amount includes the most pertinent and relevant share of the available pool.

The TREC adhoc task is representative of the analytical tasks facing today's knowledge managers, analysts and subject matter experts. Knowledge acquisition requires them to compare yesterday's knowledge inventory (acquired perceptions and analysis reports and the standards and decisions based on them) with today's knowledge input and calculate some measure of knowledge growth. They also need the means to identify current knowledge gaps and some notions of importance (what deserves attention? what is pertinent?). READWARE's knowledge types form a system of knowledge organization that is capable of meeting all these needs.

The TREC experiment shows that knowledge managers and analysts with or without subject matter expertise can use READWARE's knowledge types (concepts, topics, probes, issues and document subjects) and the advanced search features (Boolean logic, sequence enforcement, context sizing, word search, phrase search and document-level search) to identify their knowledge gaps and formulate automatically executable information requests to fulfill their information needs (READWARE queries as formalized information requests).

READWARE is efficient at TREC text retrieval because it uses a coherent unifying framework of knowledge organization. This framework is layered, well-structured, expandable and even open to integrating custom models of knowledge organization.

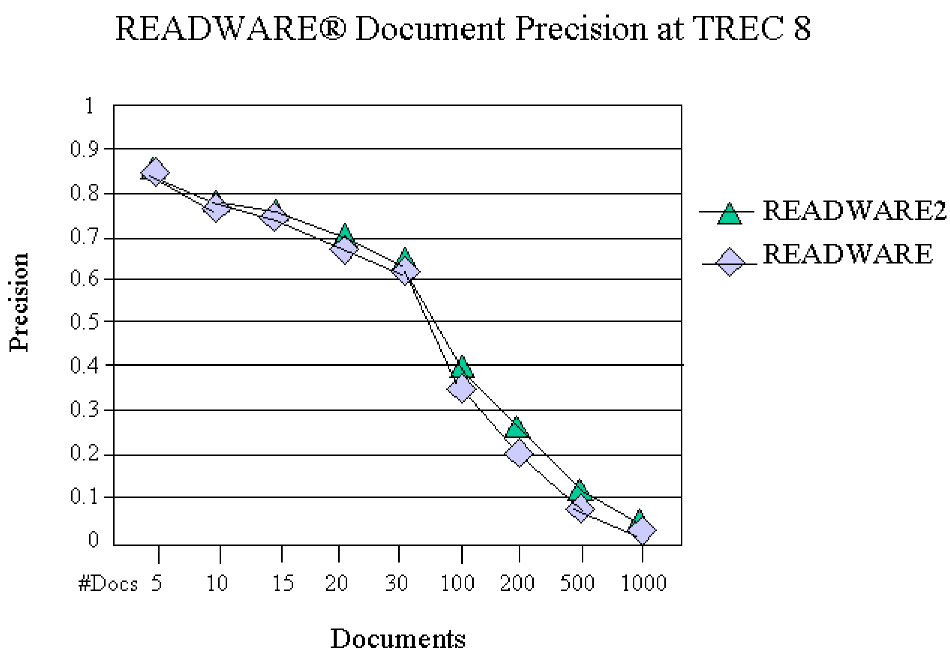

Figure 2.