Oracle at Trec8: A Lexical Approach[1]

Kavi Mahesh, Jacquelynn Kud, Paul Dixon
Oracle Corporation
500 Oracle Parkway M/S 4op10
Redwood Shores, CA 94065 USA
trec@us.oracle.com

Abstract

Oracle's system for Trec8 was the ***inter*Media Text** retrieval engine integrated with the **Oracle8*i*** database and SQL query language. interMedia Text supports a novel theme-based document retrieval capability using an extensive lexical knowledge base. Trec8 queries constructed by extracting themes from topic titles and descriptions were manually refined. Queries were simple and intuitive. Oracle's results demonstrate that knowledge-based retrieval is a viable and scalable solution for information retrieval and that statistical training and tuning on the document collection is unnecessary for good performance in Trec.

1. Introduction

Oracle's approach to Trec8 is a novel theme-based retrieval method implemented on top of traditional boolean techniques. Theme-based retrieval enables querying for documents that are *about* a certain theme or concept. The set of themes from a document together define what a document is about. Themes are extracted from documents and queries by parsing them using an extensive lexicon together with a knowledge base of concepts and relations. High precision is achieved by a disambiguation and ranking technique called *theme proving* whereby a knowledge base relation is verified in the lexical and semantic context of the text in a document.

Oracle's solution, known as interMedia Text is a part of the Oracle database with a standard SQL interface for creating indexes and issuing queries. Database integration and SQL interface make the retrieval engine highly scalable and easy to use. For Trec8 (ad hoc, manual), Oracle used its interMedia Text product without any modifications and with absolutely no training or tuning on the Trec document collection. As such, Oracle's solution is expected to deliver equally good quality and performance in other topic domains or with other document collections.

2. Theme-based Document Retrieval

interMedia Text builds an inverted index of a document collection that is either stored or referenced (via file paths, URLs, or procedures) in a database column. In addition, it also adds a hierarchy of ranked themes extracted from each document to the inverted index. Information retrieval queries are part of SQL SELECT statements and hence can be easily combined with other structured queries (e.g., to query document metadata). Queries return hitlists with relative scores in the range of 1..100 that can be used for ranking the hits.

---

1. Product availability: **Oracle8*i* *inter*Media Text Release 2 (8.1.6)** is currently available in the market. The product is also available for download for Oracle Technology Network members at http://technet.oracle.com

interMedia Text supports theme or concept-based retrieval using the ABOUT operator which extracts themes from free text queries to match against themes in the inverted index. Theme queries are inherently hierarchical in nature. For example, our query from Topic 436 (railway accidents) included 'ABOUT(rail transportation)' which is in effect equivalent to expanding the term to a large number of narrower concepts and their synonyms using the knowledge base. Figure 1 shows some of the terms in this hierarchy. However, theme queries are not expanded; instead document themes are indexed hierarchically so that documents containing any of the narrower terms would be indexed against the theme rail transportation. Hence, theme queries yield high recall numbers and run much faster than an expanded query. Recall is also increased by recognizing noun phrases and by converting synonyms and alternate forms to canonical forms.

CATEGORY: rail transportation

Conrail, Incorporated
Amtrak
Association of American Railroads
Norfolk Southern Corporation
Southern Pacific
Union Pacific Corporation
automatic train control, automatic train controls, atc system, atc systems
brakepersons, brakeman, brakemen, brakewoman, brakewomen, brakeperson
broad gauges
cog railroads
cowcatchers
gandy dancers
monorails
rail stations, railroad stations, railroad terminals, railway stations
railroad tunnels
railroads, railways
siderodromomaniacs
stationmasters
streetcars
tramways
....

SUBCATEGORY: trains

bullet trains, commuter trains, express trains, goods trains, passenger trains, shuttle trains
center dump ballast cars, chair cars, dining cars, drop-bottom dump cars, log cars, mail cars, observation cars
railcars, railway cars, sleeping cars, tank cars, tender cars
....

SUBCATEGORY: subways

London Underground
Metropolitan Railway
New York City Subway
BART - Bay Area Rapid Transit
Washington D. C. Metro
....

Fig. 1: *Knowledge base content for rail transportation. A sample of over 600 terms under this category is shown.*

Moreover, not every document that contains the narrower terms (e.g., “trains”) are indexed against rail transportation due to a process called *theme proving* which increases precision. High precision is achieved by proving document themes in the lexical and semantic context of the text in the document. Two themes prove each other if they are closely connected in the knowledge base either hierarchically or through cross references. This eliminates many bad hits arising from word sense ambiguities.

There is also no need for users to formulate extremely long queries or develop a thesaurus for a topic such as rail transportation. ABOUT queries on such terms are effective since interMedia Text has an extensive knowledge base that was built with a focus on information retrieval.

Theme queries are simple and intuitive. For example, shown below is our query for Topic 414 (Cuba, sugar, exports) which obtained the best recall and an average precision well above that topic’s median:

```
‘ABOUT(cuba) * 3, ABOUT(sugar) * 5, ABOUT(Russia) * 2,
 (ABOUT(export) or ABOUT(import))’
```

ABOUT queries can be combined with boolean operators and term weighting as shown above. The ‘,’ is an abbreviation for the ACCUMULATE operator.[1] ABOUT queries can also be combined easily with conventional text queries. For example, our query for Topic 450 (King Hussein, peace) was:

```
‘(King Hussein=Hussayn=Husayn=Husyan * 5,
  ABOUT(peace talks) * 3, peace * 2, ABOUT(Israel))
MINUS (Iraq or ABOUT(Iraq))’
```

2.1. Lexical Knowledge Base

Theme extraction depends heavily on a good quality lexical knowledge base. Oracle’s knowledge base, built in-house, contains over 200,000 concepts from very broad domains classified into 2000 major categories. These categories are organized hierarchically under six top terms: business and economics, science and technology, geography, government and military, social environment, and abstract ideas and concepts. Concept classification, choice of categories, and the hierarchical organization are all carefully designed for their usefulness in information retrieval rather than ontological purity, with a special focus on avoiding problems of semantic ambiguity. For example, unlike other linguistic or AI ontologies [Bateman et al, 1990; Bouaud et al, 1995; Guarino, 1997; Lenat and Guha, 1990; Mahesh, 1996; Miller, 1990], the top terms are not terms low in content such as “noun”, “thing”, “event”, or “role”.

The hierarchy is organized by a mix of relations such as taxonomic, partitive and other relations, the choice being driven by the salience of the relation rather than uniformity across the ontology. For example, SQL - Structured Query Language is under databases because of its strong association with databases although it is really a programming language. Other relations are encoded as cross references to avoid any ambiguity in parent-child relations. Cross references across the hierarchies play a role in theme proving to enable disambiguation and better ranking of themes.

Each concept is mapped to one or more words or phrases that are in their canonical form. Each canonical form has several alternate forms such as inflectional variants, abbreviations, acronyms, alternate spellings, and synonyms. Canonical form mappings are carefully controlled in the knowledge base to give higher

---

1. The ACCUMULATE operator is ideal for Trec queries. It behaves like OR except it assigns higher ranks to documents that match more terms and terms with higher term weights.

accuracy than a stemming approach. The total vocabulary of 450,000 terms includes a large number of proper names with an extensive coverage of names of places, organizations, and people.

Each entry has a large set of flags indicating its lexical attributes such as its part of speech as well as several attributes that indicate its information content and the degree of its importance for information retrieval. The latter set of flags has been carefully assigned to each entry in the lexicon. This is a key source of information used for recognizing important themes in a document, thereby improving theme ranking.

2.2. Theme Extraction

Known phrases are recognized using a greedy algorithm. Unknown words and proper name phrases are recognized and treated as themes. Words and phrases are normalized to their canonical forms to avoid having to expand query terms. Every normalized term is a potential theme for the document.

Precision at *n* hits is improved by theme ranking. Themes are assigned initial ranks based on their flags in the knowledge base, as well as several factors derived from the structure of the document and the frequency of the theme in the document. If a theme maps to a concept in the knowledge base and it is successfully proven in the document context, all of its ancestors in the knowledge base hierarchy are also included as themes of the document.

Ranks of parent themes are determined by their distance to the descendant themes, as well as by the accumulated strength of all descendent themes in the theme hierarchies present in the document. Theme ranks are also adjusted based on the amount of evidence from related themes that proved the theme. Documents in a hitlist are ranked by combining the scores of document themes that match query themes and any user assigned term weights in the query.

Theme proving is an efficient partial solution to the general problem of word sense ambiguity. It essentially verifies if the single predominant sense encoded in the knowledge base is valid in the current document context. Two themes prove each other if they are closely connected in the knowledge base either hierarchically or through cross references. If a theme is not proven, its parent themes are not included in the index. This eliminates many bad hits arising from word sense ambiguities.

For example, document LA082890-0005 is about a plane crash and it mentions that the plane was used to train pilots. This document is not relevant for Topic 436 (railway accidents) in spite of the words “accident” and “train”. This is included in the NIST *qrels* (i.e., it was retrieved as one of the top 100 hits for this topic by at least one system). interMedia Text did not retrieve this document at all although the word “train” is under `rail transportation` in its knowledge base. Theme proving determined that there was no other evidence for `rail transportation` in this document and hence rejected the railway sense of “train”.

3. Trec8 Run

Oracle used its Oracle8*i* *inter*Media Text Release 2 (8.1.6) product without any modifications. Only the public interface of the product was used to produce the Trec8 run. Oracle’s run was produced on a Sun Sparc Ultra 1 workstation running the Solaris 2.6 operating system.

3.1. Filtering and Indexing

Documents were extracted into separate files, one per file. An attempt was made to filter the documents to extract just the text bodies using the DTDs specified for each collection. However, since about 4000 docu-

ments did not strictly conform to the DTDs and were empty upon filtering, our final approach was to remove the LA Times index terms that were to be excluded as stated in the Trec guidelines [Trec8 Guidelines] and simply filter out all SGML tags. File names and paths were entered into an Oracle database table. A combined index (i.e., a theme index plus an inverted text index) was built on the collection thus:

```
CREATE INDEX trec7_index ON trec7_docs(doc_text) INDEXTYPE IS ConText;
```

3.2. Training and Tuning

There was absolutely no training or system tuning on the Trec document collection. The lexical knowledge base was not modified in any way. No new stoplist was created. In fact, the index was built before the Trec8 topics were available.

3.3. Query Formulation

All queries were strictly text and theme queries. No other structured criteria were used in the SQL queries. The first set of queries was constructed automatically using the natural language theme query ABOUT operator:[1]

```
SELECT doc_id FROM trec7_docs WHERE
  CONTAINS (doc_text, 'ABOUT(<title>), ABOUT(<description>)', 1)>0
  ORDER BY score(1) DESCENDING;
```

These initial free text queries were converted to preliminary boolean queries using interMedia Text's Explain feature. In effect, this amounted to extracting themes from the title and description. This involves recognizing phrases, discarding function words and useless words, and normalizing the words and phrases to their canonical forms using the knowledge base. All of the themes were accumulated in the resulting internal query (using the ACCUMULATE operator).

3.4. Manual Query Refinement

25 topics each were assigned to two members of our team. Each member browsed the top few hits from each topic to identify good and bad query terms. interMedia Text's Theme Highlighting feature was used to obtain an HTML rendering of the document with highlights. The feature highlights not only the words and phrases that were normalized to a theme in the query, but also all of its related and supporting terms (identified during the theme proving operation). This enabled the user to quickly grasp why a particular document was retrieved for the query. Figure 2 shows a relevant document for Topic 436 (railway accidents) that is highlighted for the theme query 'ABOUT(rail transportation)'.

---

1. interMedia Text assigns scores in the range 1..100. However, in the final submission, each hit had a unique weight = (1000 - rank) instead of the score in the 1..100 range. This was done to avoid any unpredictability in the results since the trec_eval program uses an unstable sorting routine (qsort), i.e., it does not preserve the ranks of hits with equal scores when computing precision and recall numbers.

NATION IN BRIEF;

MARYLAND;

FINAL CLAIM SETTLED IN FATAL **TRAIN** WRECK

From Times Staff and Wire Reports

Conrail agreed to pay $5.5 million in the last of about 370 court claims stemming from a fiery **Amtrak**-**Conrail train** wreck that killed 16 people near Baltimore in 1987.

Attorneys for Susan Schweitzer, 45, of New York, said the settlement came just before selection of a jury to hear the civil damage trial in Baltimore. The crash occurred Jan. 4, 1987, when three **Conrail locomotives** ran through a switch and into the path of an **Amtrak train**. **Conrail** engineer Ricky L. Gates later admitted that he and **brakeman** Edward Cromwell had smoked marijuana just before the crash.

Fig. 2: *HTML rendering of document LA032790-0024 (Topic 436) with theme highlighting for 'ABOUT(rail transportation)'.*

The two team members made relevance judgments of the top few hits as they browsed the documents. The resulting *qrels* was a small fraction of the NIST *qrels* but was very useful to track improvements as the users refined queries based on their findings.

Unlike NIST, our team made a three-valued judgment: YES, NO, or MAYBE. The intended semantics of MAYBE was "try to make sure this document is retrieved, but ideally it should rank below the YESs" (similar to "iffy" in [Cormack et al, 1998]). This greatly improved productivity since the members did not have to scrutinize borderline documents to make a binary YES/NO decision.

It turned out that the three-valued judgments were meaningful. Our agreement with NIST *qrels* was 86% considering only YES/NO and was 77% when MAYBEs were treated as YES. Also, NIST *qrels* included 86% of our MAYBEs. 42% of our MAYBEs were considered relevant by NIST.

3.5. Increasing Recall

Queries were manually adjusted to increase recall for some topics by substituting a concept one level above the original theme in the knowledge base. For example, for Topic 449 (antibiotics ineffectiveness), the original query had "antibiotics". However, not all drugs that are commonly considered antibiotics were under this concept. The knowledge base had a finer classification of these drugs and it appeared better to go with the parent concept of `antibiotics, antimicrobials, and antiparasitics`. The knowledge base was also used to manually expand queries by adding related terms and sibling concepts for some topics. Some terms that were only present in the narrative part of topics were also added. For example, in Topic 418 (quilts, income), "quilting books" and "quilting classes" were added from the narrative.

It may be noted that the above use of the knowledge base does not require access to the internal knowledge base structure. interMedia Text provides a Hierarchical Query Feedback feature which allows users to browse broader, narrower, and related terms for a query term in the knowledge base.

Also, many queries initially did not return 1000 hits. Although the team members found no particular reason to return more hits, the queries were expanded to return at least 1000 hits since there is no penalty in Trec for returning irrelevant documents after all the relevant ones.

3.6. Increasing Precision

Precision and relevance ranking were improved on some topics by tightening the query in several ways. Boolean operators AND and NOT were used instead of ACCUMULATE. Ambiguous terms, such as "drugs" in topic 449 (antibiotics ineffectiveness), were either deleted from the query or term weights were adjusted to lower their impact on relevance ranking. The MINUS operator (which reduces the score of a hit if the specified term is present) was also used to push borderline hits lower in the hitlist. When several terms were being ACCUMULATEd, equivalent terms were grouped under an equivalence (=) or OR clause to prevent dilution of term weights.

Some of these refinements are illustrated by our query for Topic 418 (quilts, income) which obtained the best recall and the best average precision:

'quilted * 6, {quilting books} * 5, {quilting classes} * 4, quilt=quilts * 0.1,
(ABOUT(quilts) AND (ABOUT(income) or ABOUT(fundraising))) * 2'

4. Trec8 Results

Oracle's results are summarized in Table 1. Eight topics had a 100% recall and 18 topics had over 90% recall. 5 topics had the best recall at 1000 and the best average precision. Topic 447 (Stirling engine) obtained 100% recall and 100% average precision. Table 2 shows the distribution of our per topic results against the median numbers for that topic.

Table 1: Oracle's Trec8 Results

Trec8 measure	Orcl99man
Recall at 1000	71.57% (3384/4728)
Average Precision	41.30%
R-precision	43.57%
Initial precision (at recall 0.0)	92.79%
"Final" precision (at recall 1.0)	07.91%

Table 2: Distribution of Oracle's Per Topic Results in Relation to the Median Numbers

	best	>= median	< median	worst
Recall at 1000 hits	16	19	15	0
Average precision	7	26	17	0

5. Discussion

Oracle used its manual relevance judgments to refine queries. However, relevance ranking of hits was done entirely by the interMedia Text system and was not in any way influenced by the relevance judgments. Other participants have interpreted the guidelines for the manual ad hoc track to allow a reordering of hits using manual relevance judgments. Oracle's results with such reordering are shown below for comparison.

```
Queryid (Num):       50
Total number of documents over all queries
    Retrieved:    50000
    Relevant:      4728
    Rel_ret:       3384
Interpolated Recall - Precision Averages:
    at 0.00       0.9604
    at 0.10       0.8704
    at 0.20       0.7715
    at 0.30       0.6723
    at 0.40       0.5536
    at 0.50       0.4775
    at 0.60       0.3598
    at 0.70       0.2830
    at 0.80       0.2128
    at 0.90       0.1339
    at 1.00       0.0769
Average precision (non-interpolated) for all rel docs(averaged over que-
ries)
                  0.4740
Precision:
  At    5 docs:   0.8640
  At   10 docs:   0.7940
  At   15 docs:   0.7453
  At   20 docs:   0.6930
  At   30 docs:   0.6207
  At  100 docs:   0.3740
  At  200 docs:   0.2389
  At  500 docs:   0.1214
  At 1000 docs:   0.0677
R-Precision (precision after R (= num_rel for a query) docs retrieved):
    Exact:        0.4800
```

Our results would have been better and queries simpler if we could have extended the knowledge base for Trec8 topics. Although Trec8 guidelines [Trec8 Guidelines] say that lexicons may not be extended after seeing the topics, interMedia Text provides an extensibility feature that allows any ordinary user to extend its knowledge base by importing a thesaurus in an ISO-2788 like format [ISO2788, 1986]. Yet, Oracle decided not to extend its knowledge base for Trec8.

An analysis of the few topics where our results were poor (i.e., both recall and average precision significantly below the median) indicated that in most cases, the major cause was our team members had interpreted these topics differently than the NIST evaluators. These were also the topics for which our team members had doubts and did not expect good results. For example, Topic 405 (cosmic events) depended on what was considered an event. The *qrels* indicate that the discovery of some galaxies was an event. Our team member's interpretation was that it had to be an explosion, collision, and so on in order to qualify as a cosmic event. The query had been refined to rank documents about the discovery of celestial objects

lower in the hitlist. As a result, such documents fell beyond the first 1000 in the hitlist.

It is well known that the Trec document collection and retrieval results are overanalyzed [e.g., Banks et al, 1999]. We believe that many systems, especially statistical ones are too well trained on this collection. We have demonstrated that our system can do well with no training on the collection. We are eagerly looking forward to the new document collection to be used in Trec9.

6. Conclusions

Theme queries are simple and intuitive. They are easy to explain in symbolic terms to see exactly why a document was retrieved and assigned a certain rank. Theme-based retrieval requires no training on a document set. It works well across domains and document collections. Database integration makes it scalable and easy to use. The ability to combine theme queries with boolean keyword queries and structured SQL queries offers a flexible and complete solution to information retrieval.

Knowledge-based hierarchical expansion and term normalization improve recall. Theme ranking and proving improve precision--both overall and average precision. Theme-based retrieval also finds a significant number of unique relevant documents (our submission found about 110 unique relevant documents). The ability for ordinary users to customize and extend the knowledge base makes it easy to tailor the retrieval engine to particular applications.

Oracle's results demonstrate that knowledge-based retrieval is a viable and highly scalable solution for information retrieval and that statistical training and tuning on a document collection is unnecessary for good performance.

Acknowledgments

The authors wish to thank Shamim Alpha and Garrett Kaminaga for their ideas and contributions.

References

[Banks et al, 1999] Banks, D., Over, P., and Zhang, N. Blind Men and Elephants: Six Approaches to TREC data, *Information Retrieval* 1, 7-34, 1999.

[Bateman et al, 1990] Bateman, J. A., Kasper, R. T., Moore J. D., and Whitney, R. A. A general organization of knowledge for NLP: The PENMAN upper model. Technical Report, USC/Information Sciences Institute, 1990.

[Bouaud et al, 1995] Bouaud, J., Bachimont, B., Charlet, J., and Zweigenbaum, P. Methodological Principles for Structuring an "Ontology". In *Proc. Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence* (IJCAI-95), Montreal, Canada, August 1995.

[Cormack et al, 1998] Cormack, G. V., Palmer, C. R., Biesbrouck, M. V., and Clarke, C. L. A. Deriving very short queries for high precision and recall (MultiText experiments for TREC-7). In *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC 7)*. National Institute of Standards and Technology, 1998.

[Guarino, 1997] Guarino, N. Some Ontological Principles for a Unified Top-Level Ontology. *Spring Symposium on Ontological Engineering*, AAAI Tech Report SS-97-06, American Association for Artificial Intelligence, March 1997.

[ISO2788, 1986] International Standard ISO 2788: Documentation - Guidelines for the establishment and development of monolingual thesauri, Second edition - 1986-11-15, International Organization for Standardization.

[Lenat and Guha, 1990] Lenat, D. B. and Guha, R. V. Building Large Knowledge-Based Systems. Reading, MA: Addison-Wesley, 1990.

[Mahesh, 1996] Mahesh, K. Ontology Development for Machine Translation: Ideology and Methodology. Technical Report MCCS-96-292, Computing Research Laboratory, New Mexico State University, 1996.

[Miller, 1990] Miller, G. WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4) (Special Issue).

[Trec8 Guidelines] http://trec.nist.gov/act_part/guidelines.html