Exploiting Underrepresented Query Aspects for Automatic Query Expansion

Daniel Crabtree Peter Andreae Xiaoying Gao
Victoria University of Wellington, New Zealand

TASK / PROBLEM

There is a need to refine hard queries to get good search results. Search systems can help in two ways: by automatically expanding the query with additional terms or by suggesting refinements for the user to choose between.

ABRAQ OVERVIEW

Our approach, AbraQ, has three steps:

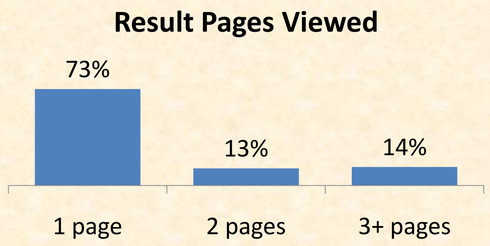
1. Identify all aspects of query.
2. Determine which aspects are underrepresented in the result set.
3. Find refinements to enhance search.

CONCLUSION

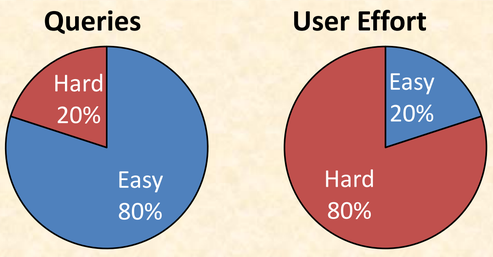
1. **Excellent Performance**: almost three-fold improvement in precision for hard queries
2. **Very efficient**: search responsiveness is unaffected and amortized average cost is only twice the original query cost
3. Provides a method of determining query difficulty
4. Requires no user involvement

INTRODUCTION

For web search, users only look at the first few documents, so precision is very important.



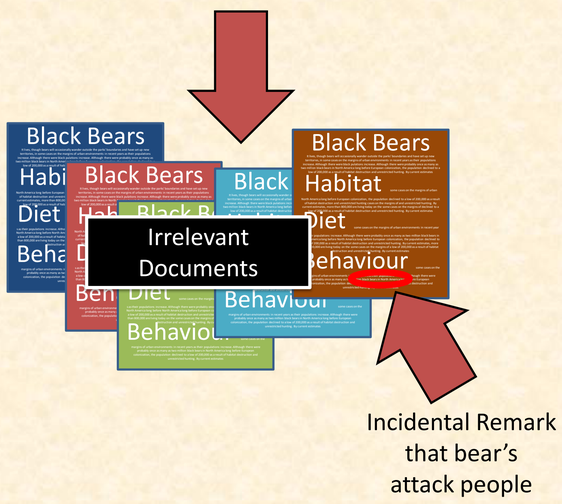
Most user time is spent refining hard queries. Hard queries have multiple query aspects.



EXAMPLE

A hard multi-aspect query where the documents with matching keywords may not primarily focus on the query's intent.

"black bear attacks people"



QUERY ASPECTS

Query Aspects are the essential components of a query.

"What flights does Air New Zealand make to Canada?"

Aspects

1. Air New Zealand
2. Flights
3. Canada

Relevant documents must represent all aspects, therefore to increase query precision, the query should be modified, such that the new query represents all aspects of the original query.

QUERY ASPECTS

AbraQ takes advantage of the fact that the order of the words within a single aspect is usually significant.

Different queries with the same intent preserve the ordering of words within an aspect. When the ordering changes, so does the meaning of the query.

- Air New Zealand Flights Canada
- Air New Zealand Canada Flights
- Flights Air New Zealand Canada
- Flights Canada Air New Zealand
- Air New Flights Zealand Canada – Non meaningful
- Air Canada Flights New Zealand – Different meaning

AbraQ identifies aspects by analyzing the relative corpus frequencies of query subsequences occurring anywhere on a page, occurring as a phrase, and occurring as a phrase in any alternate permutation.

ASPECT REPRESENTATION

As different aspects are associated with different vocabulary, AbraQ assumes that documents lacking an aspect's vocabulary do not represent that aspect.

AbraQ builds a Vocabulary Model for each aspect. The vocabulary model contains the most frequent terms that co-occur frequently with the aspect across the entire corpus. The terms are found by running sub-queries for each aspect, and each pair of aspects. Including the pairs addresses polysemous aspects.

For example, for the black bear attacks query, the vocabulary model for the "Black Bear" aspect may contain the terms "Animal", "Mammal", and "Diet", while the "Attacks" aspect may contain terms like "Fatal", "Maul", and "Danger"

If sufficient vocabulary of an aspect is present within the first N documents, then AbraQ considers that the aspect is represented.

REFINEMENT

AbraQ only attempts to refine a query when there are underrepresented aspects.

Possible refinements are constructed by combining the original query with terms from the vocabulary models of the underrepresented aspects.

Each possible refinement is evaluated by how well the documents returned from the refined query represent all the aspects.

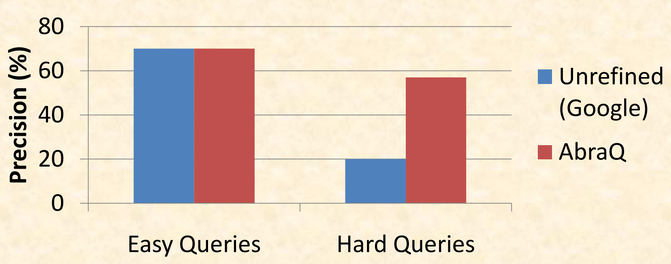
AbraQ selects the best of the possible refinements.

RESULTS

10 multi-aspect queries that were randomly selected from the TREC 2005 hard track for testing.

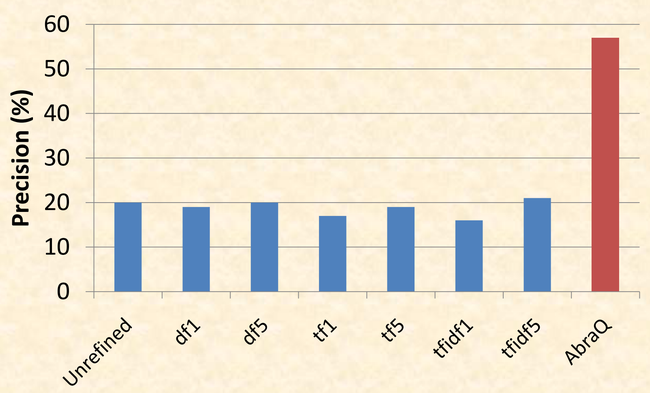
ABRAQ IDENTIFIES AND IMPROVES HARD QUERIES

Four queries were "easy" as AbraQ considered that all aspects were represented in Google's search results. The other six were "hard" and required refinement.



ABRAQ OUTPERFORMS OTHER AUTOMATIC APPROACHES

This graph shows the mean precision on the "hard" queries. On the "easy" queries, there is no significant performance difference between any of the approaches.



ABRAQ OUTPERFORMS INTERACTIVE APPROACHES

This graph shows the mean precision on the "hard" queries. The interactive approaches (Clustering, Query Log, Relevance Feedback) were given perfect user input. AbraQ User shows AbraQ's performance with user input.