# Concrete Queries in Specialized Domains:

## Known Item as Feedback for Query Formulation

Mun-Kew Leong
Institute of Systems Science
Singapore

**Abstract**

This paper investigates the use of short concrete queries as being typical of non-naive non-computer professionals in information searching and retrieval. We aim to simulate such an environment with a general corpus such as the TREC data using what we term "known item query formulation". Assuming the simulation is valid, the results suggest that such searching will satisfy the profiled user.

# 1 Introduction

We introduce the concept of a *concrete query* as opposed to an *abstract query*. We will provide a formal definition of both later. Intuitively, however, concrete queries may be thought of as those which are based on examplars, and abstract queries as those which are based on descriptions.[1] In some cases, queries can be *mixed*, i.e., a combination of the two.

## 1.1 Motivation

We approached the TREC experiments this year with an interest in the methodology of searching in specialized domains. We conducted informal discussions with non-computer professionals who, however, were well versed in using a computer, and moreover, had some to considerable experience searching for information on-line. We had the intuition that such searchers would use short queries with highly specialized terms, i.e., what we have described above as concrete queries.

### 1.1.1 Interviews

We talked separately to the following non-computer professionals:

1. a psychiatrist with more than five yrs experience using computers and who had been searching on-line and on the WWW for over three years. On-line searching experience included OPACs, Medline, etc.
2. a management consultant with a degree in Economics who had been using computers for over ten years, and who had been searching the internet for over 4 years. On-line searching experience are on business databases and the WWW.
3. a doctor with two years computer experience, and the same amount of time searching the WWW.

[1] This distinction has been otherwise termed, respectively, *referential* and *descriptive*.

The aim of the discussion was to elicit the form of the queries which these users employed to search their own areas of expertise on-line, and to get an idea from them if the way they searched on-line was similar to how their colleagues did.

While this survey was admittedly informal, and there were only three persons polled, the following information was noted:

- the users mainly used keyword (non-Boolean) search
- the query terms used depended on the search. Names of procedures, drugs, people, companies, etc., were often used. Descriptions of the things to be searched were rarely written out before hand, except as reminders.[2]
- all three users knew of, and sometimes used, and's and or's in their queries. None of them used not's. If they used Boolean operators, it was usually and's, and only when there were too many records found, and they could not find relevant records/documents among them.
- the users all revised their queries based on results of previous searches.
- queries varied in length. Initial queries were short, but revised queries could be quite long
- only one of them knew about phrase operators and mandatory operators on WWW search engines.
- all of them were satisfied with their ability to search. They thought that they could find the information they looked for most of the time.
- none of them knew how their colleagues searched. They thought it would be similar.

While it may not be clear in the observations above, all three users agreed that short queries using names and a few specialized terms would give good results.

### 1.2 Aim

We therefore conceived an experiment with the following aims:

1. determine if concrete queries would actually give good results
2. determine if manually constructed concrete queries would perform better than relevance feedback using a document which contained the same concrete search terms
3. determine if we could meaningfully use the TREC structure to simulate testing of concrete queries. This would allow us to continue experiments with concrete queries while taking advantage of the TREC topics and relevance judgements.

## 2 Concrete and Abstract

Taking an example from the TREC topic 339, an abstract query would resemble:

> drugs being used in the treatment of Alzheimer's Disease

whereas a concrete query would be something like the following:

[2] Interestingly, the reminders were not descriptions of the information to be looked for, but rather the problem to which the information was to be applied.

tacrine hydergine velnacrine selegiline THA cognex mentane

Obviously, the latter (concrete) query requires a much more specialized knowledge of the domain queried. That, however, is what we believe are the types of queries used by sophisticated searchers (doctors, scientists, etc.) and which we are interested in exploring further.

### 2.1 Formal Definitions

This section lays a preliminary foundation for formal work on concrete and abstract representations, but a more complete exposition is beyond the scope of this paper.

We define a *term* as a non-empty sequence of characters (in any orthographic system, e.g., the Roman alphabet, or the Chinese character set).

We define a *representation scheme* as the set of terms used to index or represent a domain (corpus) of objects (documents). For example, a representation scheme could be an inverted word index or the set of features in a vector model.

Formally, given a domain of objects to be represented, a representation scheme is *concrete* if it partitions the domain into **non-overlapping** classes of objects. For example, if a set of documents was partitioned by the terms "round" and "square", no document could be in both partitions since there are no round square objects. The Vienna Classification for trademarks would be an example of a concrete representation scheme.

Conversely, the representation scheme is *abstract* if it partitions the domain into **possibly overlapping** classes of objects. For example, suppose an indexing scheme partitions a corpus of documents based on the two terms "blue boxes" and "rectangular boxes", and there are documents which talk about rectangular blue boxes, and are thus members of both classes, then that indexing scheme (or representation scheme) is abstract.

We define a term as *concrete* if it is an element in a concrete representation scheme, and *abstract* if it is an element in an abstract representation scheme. We can informally think of *concrete* and *abstract* as properties of terms used to index or represent objects (documents) over a given domain (corpus).

In reality, however, a corpus of documents would have both a concrete and an abstract representation scheme. The terms used in each scheme are, unfortunately, indistinguishable in and of themselves (as strings of characters). Nevertheless, relative to a given corpus, we can still think of terms as being abstract or concrete.

## 3 The Experiment: Modelling Concrete Queries

We attempt to emulate this scenario in the generalized TREC environment using what we term "known item query formulation". This is a manual 2-step query formulation method.

### 3.1 Terminology

The following terms are used in the description of the experiment, which may require explanation:

- **highly relevant:** there is no distinction in degree of relevance in the TREC relevance judgements, but the subject was instructed to use a document (or part of a document) which addressed the topic and which contained concrete terms such as names (of people, places, procedures), objects, etc. Such a document was termed *highly relevant*. A document is *relevant* if it addressed the topic but did not contain concrete terms.
- **topic:** a topic is the TREC topic.
- **query:** a query is an instantation of the TREC topic as fed to the search engine.
- **iterative construction:** a query is constructed and fed into the search engine. The subject makes use of the results of the query to amend his or her earlier query. This is done until the query satisfies the required stop condition, or the subject runs out of time.

### 3.2 Method

The latest TREC ad-hoc topics (topics 301 to 350) were used to establish the search task for the subject. For each TREC topic,

1. the subject was given 20 minutes to find one or more documents deemed highly relevant (see above) in the corpus. Tools used were grep and its variants. There was no access to the search engine in this step.

   It is possible that the subject not find a highly relevant or even a relevant document within the time limit.

   The subject was asked to keep note of the documents deemed (highly) relevant. If none were found, then this was to be stated.

2. the subject was then given 10 minutes to iteratively construct a query which returns at least one of the documents found in Step 1 in the top 10 ranked documents using the search engine. The subject was instructed to construct short queries if possible, and to stop once the condition was satisfied.

   This query is saved and used for the run. No query expansion or pseudo-relevance feedback was employed.

   It is possible that the subject not be able to construct a query which returns any relevant documents in the top 10 ranked list.

   In the case that the subject not find any relevant documents in Step 1, the subject was asked to construct what he or she thought to be the best query.

Step 1 basically creates the scenario that the subject has a "resource" (in this case the documents deemed relevant, in the true case, their domain expertise) from which they may mine relevant concrete terms from which to construct their query. The time limit was imposed purely to make the experiment manageble.

Step 1 may also be thought of as identifying a *known item* for retrieval. Once that item is found in the top 10 documents in Step 2, the query is fixed. This was what was meant by the sub-title "known item as feedback for query formulation".

Just to be clear, in Step 2, the subject has access to all and only the following in the construction of his or her query:

- the original TREC topic statement in its entirety
- documents found in Step 1 which are deemed highly relevant, or which are deemed relevant.
- documents returned by the search engine in Step 2, their order, and their contents.
- any personal knowledge with which to construct or augment the queries

# 4 Running the Experiment

## 4.1 Subject and Experimental Conditions

There are 50 TREC topics, and with half an hour for each topic, this is a fairly fatiguing and time consuming experiment. Thus, so as not to stress the subject (with consequent effects on performance), and to minimize fatigue, the experiment was conducted over the course of one week. Because of time constraints, only one subject was used. Estimated total time of the experiment was 22 hours. Actual time spent logged on to the workstation was just over 18 hours.

The subject was a staff of ISS, with about 10 years of computer experience, and was well versed in using a UNIX workstation, and with using UNIX system and search tools such as grep, fgrep, cat, more, etc. The subject was allowed to use (and did use) multiple windows, running several searches in parallel.

### 4.1.1 Problem with Relevant Documents

There was a problem in the experiment. The subject did not keep a list of the documents deemed relevant in Step 1 for each topic. A list was reconstructed by the subject the following week listing at least the topics for which relevant documents were found, and for which relevant documents were not found. Where relevant documents were found, the subject (using the search engine) tried to identify what he had used as relevant and highly relevant documents for each topic. He is sure there are missing relevant documents in his list, but is quite sure that no documents were falsely identified as being relevant. Recall again that "relevant" in this context is with respect to Step 1 of the experiment, and not with respect to TREC relevance judgements.

## 4.2 Example queries

We examine some of the constructed queries, where relevant documents were found in Step 1 and where no relevant documents were found. We will later analyse the results along the same lines.

### 4.2.1 Topics with relevant documents

Out of the 50 topics used in the experiment, the subject found highly relevant documents for 29 of them in Step 1. The following two examples particularly show the addition of concrete terms.

- **Topic 331:** World Bank Criticism

> **Desc:** What criticisms have been made of World Bank policies, activities or personnel
> **Narr:** This query is looking for any instances where the World Bank has been accused of things like not being responsive to the unique problems of individual countries, of being too strict in its policies, of pursuing agendas that are biased because of their benefits to western countries, of being no longer useful or practical, of its personnel being difficult to work with, etc.

**Constructed Query:** world bank criticism criticize environment funding damage lending projects barber conable lewis preston

- **Topic 339:** Alzheimer's Drug Treatment

> **Desc:** What drugs are being used in the treatment of Alzheimer's Disease and how successful are they?
> **Narr:** A relevant document should name a drug used in the treatment of Alzheimer's Disease and also its manufacturer, and should give some indication of the drug's success or failure.

**Constructed Query:** alzheimer tacrine hydergine velnacrine selegiline THA cognex mentane

However, other queries where highly relevant documents were found did not seem to be very concrete:

- **Topic 307:** New Hydroelectric Projects

  **Constructed Query:** new hydroelectric projects construction

- **Topic 312:** Hydroponics

  **Constructed Query:** hydroponics

As it turns out, these simple constructions were sufficient to retrieve the relevant documents (from Step 1) in the top 10, and so the query was frozen at that point.

#### 4.2.2 Topics with no relevant documents

The 21 topics with no relevant documents found in Step 1 were examined and none of the constructed queries had concrete terms. Examples, however, varied widely, with some examples as follows:

- **Topic 335:** Adoptive Biological Parents

  **Constructed Query:** stepparents stepfather stepmother stepchildren stepfamilies biological parent court decision

- **Topic 319:** New Fuel Sources

  **Constructed Query:** fuel research

### 4.2.3 Query Length

One of the issues with query formulation has been query length. The average query length reported for Web search engines[3] is between two and three words long. The average length of the TREC topics, using only the Title and Description fields is 20.2 words long.

Based on the queries constructed above,

- the mean length of all the queries was 7.1 words
- the mean length of queries where relevant documents were found was 5.7 words
- the mean length of queries where no relevant documents were found was 4.2 words.

As we said in our observations in Section 1.1.1, queries varied in length, and started out short, but revised queries could be quite long. This is not disproved by the above evidence.

# 5 Retrieval Results

We present the general results from running the ad-hoc task both to situate the engine with respect to other systems, and to provide baselines for interpretation. Again, we stress that no pseudo-relevance feedback was employed. In our engine, this typically (based on previous TREC topics and data sets) increases average precision by about 10%.

## 5.1 General results

We are interested in two of the submitted ad-hoc runs, that of *iss97man* which is the manual query formulation (i.e., the experiment), and *iss97s* which is the baseline formed from the short query task, i.e., the ad-hoc task run using only the Title and Description fields from the topics. These results are shown in Figure 1. As expected, manual query

| | iss97man | iss97s |
|---|---|---|
| Recall/4611 | 2792 | 1599 |
| Precision @0.2 | 0.4512 | 0.1889 |
| Precision @0.5 | 0.2517 | 0.1206 |
| Average Precision | 0.2576 | 0.1135 |

Figure 1: Ad-hoc results for *iss97man* manual query formulation

formulation produces much better results than the automatic short queries, especially with no pseudo-relevance feedback.

Incidentally, to provide perspective, we compare the performance of the manual run, *iss97man* relative to all the manual runs submitted for TREC-6. For relevant documents retrieved, *iss97man* performed ≥ median on 42 of the 50 topics, with 8 matching the best

[3] Doug Cutting, Excite, during *Panel Session on "Real World" Information Retrieval*, SIGIR'97

results for a topic. For average precision, it performed $\geq$ median on 35 of the topics, however with only 1 matching the best results for a topic.

### 5.2 Investigation results

The number of topics in which highly relevant documents (in the sense used in this experiment) were found was 29. This means there were 21 documents where no such documents were found. The topics can therefore be divided into two classes: with-relevant and without-relevant. Since all the queries were constructed by the same subject, the without-relevant class is suitable as the *control* topics. The with-relevant class is the *experimental* topics.

The actual results for the two classes are shown in Figure 2. Since the number of topics

| | experimental (29 topics) | control (21 topics) |
|---|---|---|
| Recall | 73% (1759/2401) | 47% (1033/2210) |
| Precision @0.2 | 0.5348 | 0.3357 |
| Precision @0.5 | 0.3238 | 0.1520 |
| Average Precision | 0.3221 | 0.1685 |

Figure 2: Raw Control and Experimental results

in each class is different, where appropriate (as for the Recall figures), comparisons between the control and experimental topics will be done on a percentage basis. The difference between the two are obviously significant.

Comparing the performance of the control and experimental topics relative to the median results of all the manual runs submitted for TREC-6, we have Figure 3 for Recall performance, and Figure 4 for Average Precision peformance. We include percentage fig-

| | experimental (29 topics) | control (21 topics) |
|---|---|---|
| matching top | 7 (24%) | 1 (5%) |
| median or better | 25 (86%) | 17 (81%) |
| less than median | 4 (14%) | 4 (19%) |

Figure 3: Recall median performance for Control and Experimental Topics

ures for easier comparisons. Numbers do not add to 100% because "matching top" figures are included in "median or better" figures.

Looking at Figure 3, we see that recall figures are not particularly different, a 5% greater "median or better" topics for the experimental topics not being significant with such a small sample.

Looking at Figure 4, we see that the average precision figures are very different. Both the "median or better" figure and the "less than median" differ in consistent (i.e., opposite)

| | experimental (29 topics) | control (21 topics) |
|---|---|---|
| matching top | 1 (3%) | 0 (0%) |
| median or better | 24 (83%) | 11 (52%) |
| less than median | 5 (17%) | 10 (48%) |

Figure 4: Precision median performance for Control and Experimental Topics

directions by about 30% in both cases. This is significant.

## 6 Conclusion

We look at Section 1.2 where we detailed our experimental motivation. We have confirmed that concrete queries do give better results, specifically with higher precision. Due to lack of time, we were unable to confirm if manually constructed concrete queries perform better than relevance feedback using documents containing the same concrete search terms. We have also determined that it is possible to use the TREC structure to simulate testing of concrete queries.

As we said, the evidence supports our intuition that expert users tend to retrieve more precise results. We believe this is based on their specialized and concrete vocabulary, but this will have to be investigated further.

While it is good to have the hypothesis experimentally confirmed, even with such a small sample, there is really nothing particularly unexpected about the results. There are, however, two other interesting thrusts arising from this experiment. The first is the formal foundational approach in Section 2, which needs to be developed further. The second is whether the experimental procedure outlined in this experiment can be automated, i.e., if known item feedback can be automatically invoked in non-interactive query formulation.