Interactive Document Retrieval Using TOPIC®
(A report on the TREC–3 experiment)

Richard M. Tong
Verity, Inc.
1550 Plymouth Street, Mountain View, CA 94043

1 Introduction

This paper contains a description of the experiments performed by Verity, Inc. as part of the Third Text Retrieval Conference (TREC-3).[1]

Verity participated as an Interactive Category A system and performed the full set of routing and adhoc experiments, submitting complete sets of results for both experiments.

Section 2 of the paper contains a review of the TOPIC system itself and the data structures it produces. Section 3 of the paper contains a description of our experimental procedure. Section 4 contains an analysis of our official results. Section 5 contains some general comments on overall performance and a brief discussion of possible future directions. The Appendix contains additional details of our procedures, as well as an analysis of topic 122, the interactive "focus topic."

2 TOPIC System Configuration for TREC-3

TOPIC® is a commercial full-text retrieval system that is available for a wide range of hardware and software environments. For the TREC-3 experiments we used version 4.0 of the TOPIC client running on either Sun SPARC 2 class machines with an X/Motif GUI, or on 386/486 class machines running Windows. The TREC data and indexes themselves were stored on a Sun SPARC 2 file-server with approximately 4.6 Gbyte of disk storage. This file sever was used exclusively for TREC during the period of the experiments.

We used the standard set of Verity database administration tools to build the TREC database. We chose to extract only the document ID (i.e., the <DOCNO> field) and a "title" (e.g., the <HEAD> field in the AP collections, the <HL> field in the WSJ collections) for the structured elements in our databases, and then treated each document as a regular full text document. We did not use a stop word list, although we did remove the SGML tags, and so we chose to create just an inverted index on word positions within documents (as opposed to an index that also recorded sentence and paragraph positions) to keep the index at a reasonable size. The final index (for all three disks) was approximately 1.5Gbyte and took approximately 112 hours to build.

We used the training data provided for disks 1&2 with respect to topics 101–150 to define a series of "ground truth" datasets that we modelled as Topics. These ground-truth Topics were made available to all searchers as part of the interactive routing query development process. We also used TOPIC's launch facility to allow searchers to access the official TREC scoring program so that the performance of the routing queries (with respect to the training data) could be assessed as they were being developed.

3 The TREC-3 Experiments

For TREC-3 we participated in Category A and produced two sets of results in each of the routing and adhoc experiments. These results sets are labelled, naturally enough, TOPIC1, TOPIC2, TOPIC3 and TOPIC4.

TOPIC1 and TOPIC2 are the routing results (i.e., the results for topics 101 through 150) — TOPIC1 corresponding to the initial, manually generated, set of queries made available to all searchers participating in the experiment, and TOPIC2 corresponding to the final routing queries developed by the searchers. TOPIC3 and TOPIC4 are the adhoc results (i.e., the results for topics 151 through 200) — TOPIC3 corresponding to a set of automatically generated queries that we made available to all searchers, and TOPIC4 corresponding to the final adhoc queries developed by the searchers. TOPIC2 and TOPIC4 are thus our "official" submissions. This information is summarized in Table 1.

1. Requests for further information about the TREC-3 experiments, or Verity's line of information retrieval products, should be directed to the author either at the address above or electronically to `rtong@verity.com`.

Table 1: Results Identification

Result Set	TREC Category	Method	Queries
TOPIC1	Routing A	Manual	Initial set of queries; based on Verity's adhoc queries from TREC-2 [1].
TOPIC2	Routing A	Interactive	Final set of queries; the official routing queries for TREC-3.
TOPIC3	Adhoc A	Automatic	An automatically generated set of queries; based on a simple transformation of the TREC information need statements.
TOPIC4	Adhoc A	Interactive	Final set of queries; the official adhoc queries for TREC-3.

Since our approach to developing queries is based on the construction of conceptual representations of the domain of the queries, the final queries for both the routing and adhoc experiments are generally the result of several interactions between the developer of the query and the test collections. Our approach is thus fundamentally interactive, and for that reason we elected to participate in TREC-3 as an interactive system.

3.1 Experimental Procedure

The experimental procedure for Verity's TREC-3 experiments is briefly summarized here. It consisted of the following main steps:

- **Training of searchers.** All searchers participated in a short training session in which the basic TREC experiments were described and the test and evaluation procedures discussed. The novice searchers were also given a short tutorial introduction to TOPIC and query building. Those who needed it were given personal instruction in the use of TOPIC for the TREC experiment. Searchers were not, however, given any specific guidelines as to how exactly they should develop queries (i.e., no attempt was made to impose a standard procedure or a Verity "style"); they were simply instructed to make use of whatever features of the TOPIC query building language they found most useful.
- **Interactive routing query development.** All searchers had access to the ground-truth Topics, the general purpose thesaurus shipped as part of the standard TOPIC product, and the initial set of routing queries based on Verity's TREC-2 adhoc queries. Each searcher was asked to "sign up" for one of more topics and was then free to work on the queries to the extent that their individual schedules permitted. No constraints were placed on the amount of time searchers could (or should) spend developing queries, nor were any constraints placed on the number of "iterations" that needed to be performed. Searchers were asked to keep notes on progress in query construction. Query development ceased on a specific date (defined as one week before the routing results were due at NIST), at which time searchers submitted whatever they had developed up to that point.
- **Interactive adhoc query development.** The only assistance provided to searchers in the adhoc experiment was the set of automatically generated adhoc queries and the standard thesaurus. Searchers were given a fixed amount of time to develop queries corresponding to the new adhoc topics, and were asked to keep notes on their query construction efforts. Again, no special constraints were imposed on the process by which queries should be developed, and searchers submitted whatever they had developed by the end of the time period.
- **Generation of test data.** Once the query development phase was completed, the individual queries were merged to form the "official" query set. We then used an interactive, command-line version of TOPIC, called VSH, to generate the official results sets.[2]

3.2 Searcher Characteristics

Twelve Verity staff members participated in the TREC experiments as searchers. They ranged widely in TOPIC experience and job function, and they developed different

2. We discovered after the fact that this occasionally results in official scores that are different from the ones computed from within TOPIC v4.0 (i.e., the version of the product used by the searchers). This is due to the *order* in which results are generated within a relevance bracket. VSH generates them in the reverse order from TOPIC v4.0, so that if there are many documents in the relevance bracket at which the 1000 document cut-off occurs, it is possible that relevant documents could be above the cut-off in the TOPIC v4.0 results list, but below the cut-off in the VSH results list. This usually makes only a small difference, but occasionally can have significant impact on the recall scores.

numbers of queries (both routing and adhoc) as their schedules permitted. Summaries of these characteristics are given in Table 2. The column labels #R and #A denote, respectively, the number of routing and adhoc queries developed by each searcher. The designation as "expert" or "novice" refers to experience with TOPIC the product, and does not necessarily imply that the searchers were expert searchers, although several were. None of the searchers had any special expertise with respect to the subject domains of the routing and the adhoc topics.

Table 2: Searcher Characteristics

ID No.	Experience Level	Job Function	#R	#A
01	Expert	Engineering	23	29
02	Expert	Engineering	–	1
03	Expert	Education	1	–
04	Expert	Sales	5	5
05	Expert	Tech. Support	–	5
06	Novice	Engineering	1	–
07	Novice	Engineering	3	4
08	Novice	QA	2	3
09	Novice	QA	1	1
10	Novice	QA	–	1
11	Novice	Education	1	–
12	Novice	Documentation	–	1

4 Discussion of Official Results

This section contains an analysis of the official results generated by NIST. Section 4.1 discusses the results from the routing experiment, Section 4.2 discusses the results from the adhoc experiment.

4.1 The Routing Experiment

Of the 50 queries submitted for the official routing experiment, only 36 were actually the result of interactive development. Of these 36, 29 were developed by TOPIC experts and 7 by TOPIC novices. The remaining 14 were just the initial manual queries.

An analysis of the notes kept by searchers suggests that the average time spent developing queries varied widely, from as little as 30 minutes to as much as 20 hours. On average though, experts spent significantly less time than novices on query development; approximately 90 minutes per query versus approximately 400 minutes per query.[3]

A comparison of the initial and final queries shows that most changes were relatively minor, in that they typically involved adding specific query terms and adjusting weights and operators. However, in approximately 10 queries there was more significant changes, including complete rewrites and major structural modifications.

The results generated by the two sets of queries with respect to the *training* data (designated TOPIC0 — corresponding to the initial queries, and TOPIC00 — corresponding to the final queries) indicate a generally small, but positive, improvement in scores. Table 3 shows the main performance measures. Inspection of the results on a per-query basis do not reveal any significant patterns of behavior *vis a vis* these measures, and so we are not able to draw any particular conclusions about the query building strategies of searchers. Anecdotal evidence suggests that most searchers adopted a strategy that amounted to getting the Rel_Ret number to a "reasonable" level and then focusing on the Prec. @N measures.

Table 3: Routing Queries on Training Data

Performance Measure	TOPIC0	TOPIC00	Δ%
Rel_Ret	6599	7243	+9.76
Av_Prec.	0.2510	0.2913	+16.06
Prec. @10	0.5840	0.6160	+5.48
Prec. @100	0.4582	0.4766	+4.02
R-Prec.	0.3174	0.3459	+8.98

Table 4 shows the same performance measures for the official routing test data. Notice that the revised queries again outperformed the initial queries. The improvements in performance are, however, significantly larger than on the

3. The times are very approximate, based as they are on the notes kept by the searchers. For the novices, the time spent often reflects lack of familiarity with both the product and the general problem of searching large text collections. No searcher kept detailed enough records that we can say anything definitive about the "number of iterations" or the "number of documents read" during the query building process, or about the features of the product that users found especially useful or frustrating.

Table 4: Routing Queries on Test Data

Performance Measure	TOPIC1	TOPIC2	Δ%
Rel_Ret	5371	5952	+10.82
Av_Prec.	0.2243	0.2774	+23.67
Prec. @10	0.4740	0.5280	+11.39
Prec. @100	0.3382	0.3880	+14.73
R-Prec.	0.2786	0.3343	+19.99

training data. The absolute level of the performance is less than on the training data (as one would expect) — although perhaps not significantly so; Av_Prec. is reduced by 4.77% and R-Prec. is reduced by 3.35% when we compare TOPIC00 to TOPIC2.

Examination of the results on a per-query basis shows a strong correlation between performance improvements on the test data and improvements on the training data. That is, an increase in the performance of a query on the training data usually corresponded to a similar increase on the test data, although not always to the same degree. This suggests that the changes made by the searchers were not in response to any specific characteristics of the training data, but really did reflect an improved understanding of the intent of the queries.[4]

Comparison of our results with the summary results for Category A systems, shows that most of the TOPIC2 queries were at the median or somewhat below on the average precision measure. Of the rest, two were complete failures (that is, they returned no relevant documents), five were close to the minimum, and three posted the maximum average precision score. Figure 1 shows the differences between the TOPIC2 results and the medians for average precision.

Preliminary analysis indicates little correlation between performance and searcher expertise or job function. In fact, of the two queries that were complete failures (for topics 131 and 141), one was done by an expert and one by a novice; and of the three queries that did best in category A (for topics 105, 111, and 145) two were done by experts, and one was done by a novice.

We also did not find any interesting correlation between performance and the "hardness" of topics 101–150. In her RIAO'94 paper [2], Harman defines hardness in terms of the relative recall, which is a measure designed to show how systems perform at the early stage of retrieval, and while this appears to be a reasonable predictor of the median average precision scores for the TREC-3 routing experiment, we found it to be less effective in predicting our own average precision scores.

The overall results for the routing experiment are somewhat surprising since we had predicted that we would achieve significance performance gains *vis a vis* the TREC-2 results. Our tentative explanation is that while our interactive approach can be made to work very well indeed (*vide* the performance on topics 111 and 145), human searchers are not, in general, very effective at exploiting large amounts of training data compared with the statistical techniques used by the majority of the automatic methods participating in TREC-3.

4.2 The Adhoc Experiment

All 50 of the adhoc queries for the TOPIC4 results set were generated interactively by the searchers, and of these, 40 were generated by TOPIC experts and 10 by novices. As in the routing experiment, the average time spent by novices and experts in developing queries was markedly different; approximately 80 minutes per query for experts versus approximately 274 minutes per query for novices. While this difference is of a similar magnitude as in the routing experiment, the much reduced novice times are at least partly explained by the fact that those novices who took part in both experiments had, by the time of the adhoc experiment, achieved some skill with TOPIC, thus enabling them to develop queries more efficiently.

The anecdotal evidence from the notes kept by searchers suggests that the nature of the adhoc test also acted to reduce the amount of time spent on each query. That is, since there is no "ground truth" against which to evaluate the results of the query building effort, searchers often seemed satisfied, or at least were not willing to spend any further effort, after relatively little interaction.[5]

The searchers in the adhoc experiment were not constrained in any way with respect to the number of iterations they could perform, the number of documents they could view, the number of times they could view any given document, or in any other aspect of their interactions with the

4. As of this writing, we have not performed a detailed statistical examination of these correlations, so these observations should be treated with some caution.

5. Again, the searchers' notes are not detailed enough to determine specific figures for the number of iterations or the number of documents read. An examination of the final adhoc queries shows, however, that many of them are very simple (both in terms of the number of terms they include and the degree of structuring they exhibit), which supports the conjecture that little effort was spent developing them.

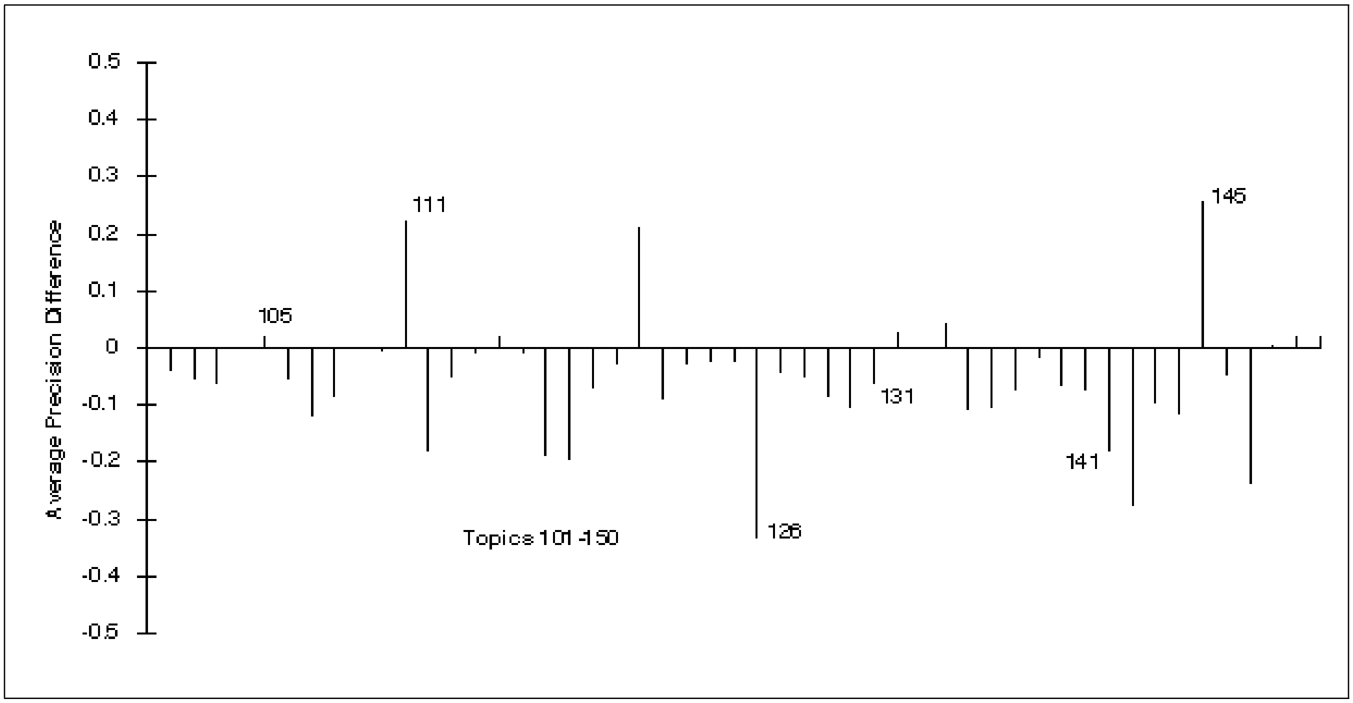

Figure 1: Average Precision Difference (Topics 101–150)

system. This is very much a "free style" approach to query development, in contrast with more formal methodologies such as the various form of frozen rank testing, but is strongly representative of how our system is used in practice. The results of the adhoc experiment (i.e., the TOPIC4 results set) should therefore be viewed as an instance of how a manual, interactive system can perform, rather than as a set of data that is directly comparable with, say, the automatic systems participating in the adhoc experiment.

The overall results for the adhoc experiment are shown in Figure 2. Notice that compared to the median average precision scores for all Category A systems the results look quite good. There is much more variability than in the routing experiment, but on balance the scores are above median, with five of the queries (for topics 158, 159, 172, 181 and 183) posting the best average precision score. As with the routing experiment though, we do not find any significant correlation between searcher experience and query performance. Somewhat counterintuitively, however, the results do seem to be generally lower (relative to the median) and more variable as the median scores increase. We have no immediate explanation for this.

5 Commentary

Verity's approach to full-text information retrieval is based on a belief that users need to be provided with a query language that enables them to built personalized semantic models of their information needs. Within this context, we see query development as a form of conceptual modelling, and thus have designed the TOPIC system to be fundamentally interactive. Although we provide a number of automatic query expansion methods, our philosophy is that users need to be actively engaged in the query development process if they are to be convinced of both the soundness of the system and the quality of the search results.

The results of the TREC-3 experiment show that this approach can be very effective, especially in situations where the searcher cannot draw upon large numbers of exemplars. However, the TREC-3 results also show that automatic systems that are able to extract significant search terms and phrases from large sets of training data performed better, on average, than the manually generated TOPIC queries. We should probably not be surprised by this. Humans are not particularly well suited to the task of determining statistical correlations among large sets of data, so that to the extent that good retrieval performance depends on finding those correlations, automatic methods will outperform manual ones. We note, though, that even in the routing experiment, the TOPIC queries outperformed all others in three cases. This suggests that human expertise and judgement can be the critical factor in developing effective queries. Further analysis of our queries, as well as a comparison with those generated by automatic systems, will of course be required to verify this conjecture.

The experimental data we collected with respect to the performance of individual searchers was, unfortunately, not particularly informative. The only definitive statements we

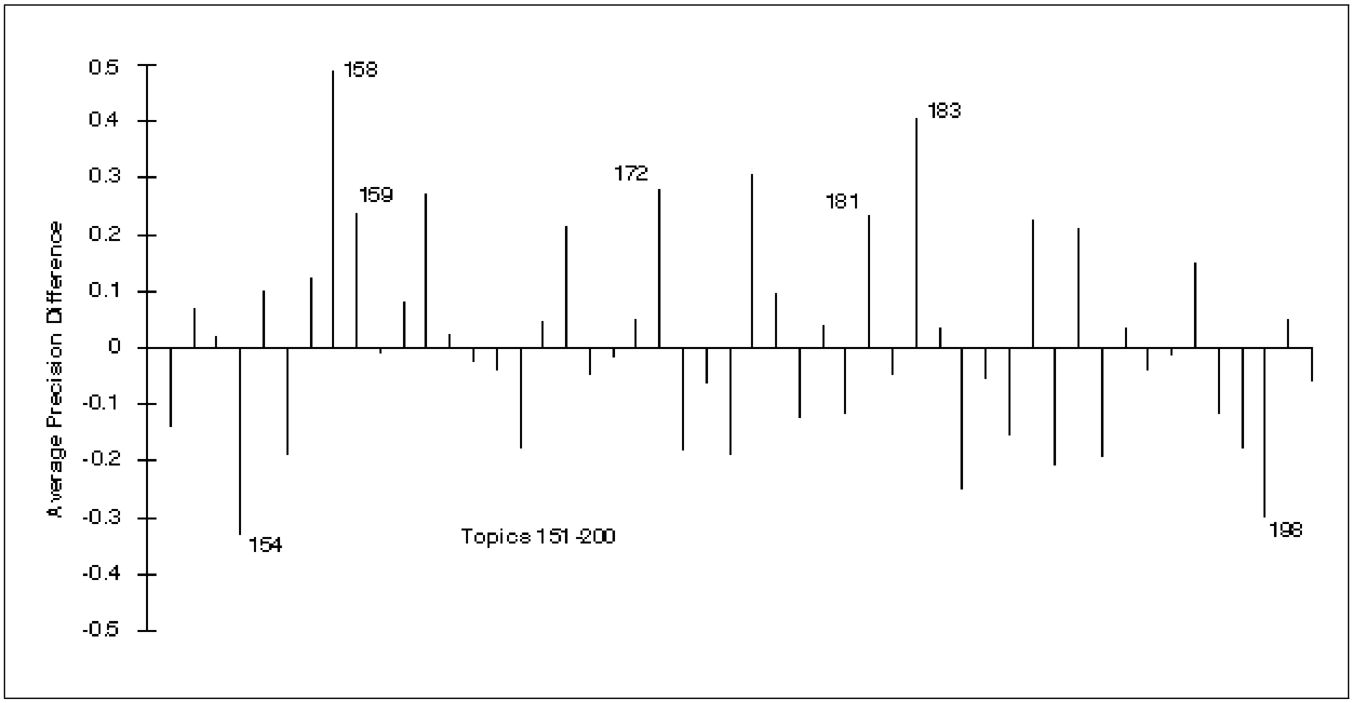

Figure 2: Average Precision Differences (Topics 151–200)

can make are that "experts" spent less time developing queries on average than "novices," and that less time on average was spent on the adhoc queries than on the routing queries. We found no significant correlations between performance and user expertise, nor between performance and query "hardness." Neither did we find any correlation between performance on individual queries and the amount of time spent developing them. In the case of the routing experiment, we did find that the performance improvements realized on the training data were also realized on the test data, which suggests to us that the searchers were indeed building conceptual models, rather than just responding to surface level characteristics of the training data.

Overall, we feel that this was a worthwhile exercise for us. While our results on the routing experiment were not as strong as we had anticipated, we did observe some searcher behaviors with respect to our product that will allow us to make improvements to the Topic construction process. The adhoc results were very satisfactory, despite the relatively small amounts of effort most queries received, and show that the TOPIC query modelling language can be used to great effect.

References

1. Lehman, J. W., Reid, C. A., et al. Knowledge-based searching with TOPIC. In: Harman, D. K., ed. The Second Text Retrieval Conference (TREC-2). NIST Special Publication 500-215, Gaithersburg, MD: March 1994.
2. Harman, D. Analysis of Data from the Second Text REtrieval Conference (TREC-2). Proc. RIAO'94: Intelligent Multimedia Information Retrieval Systems and Management, pp 699–709. Rockefeller Univ., New York, NY. October 11–13, 1994.

A Appendix

This appendix contains some additional details of the experimental procedures used in the routing experiment. In particular, it contains a description of the TOPIC user interface together with a description of the main functional elements in that interface, and a detailed analysis of the results for topic 122.

A.1 The TOPIC User Interface

TOPIC is a commercial full-text retrieval system available for a wide range of hardware and software platforms, and can be configured in a number of ways for different environments. A typical query development screen from the X/Motif version of the user interface is shown in Figure 3.

The display is divided into two main sections. The upper portion contains the query (in this case a partially expanded Topic tree for topic 146), and the lower portion contains the hit list. The relative sizes of these two panes are completely controllable by the user, as is, of course, the overall geometry of the TOPIC window.

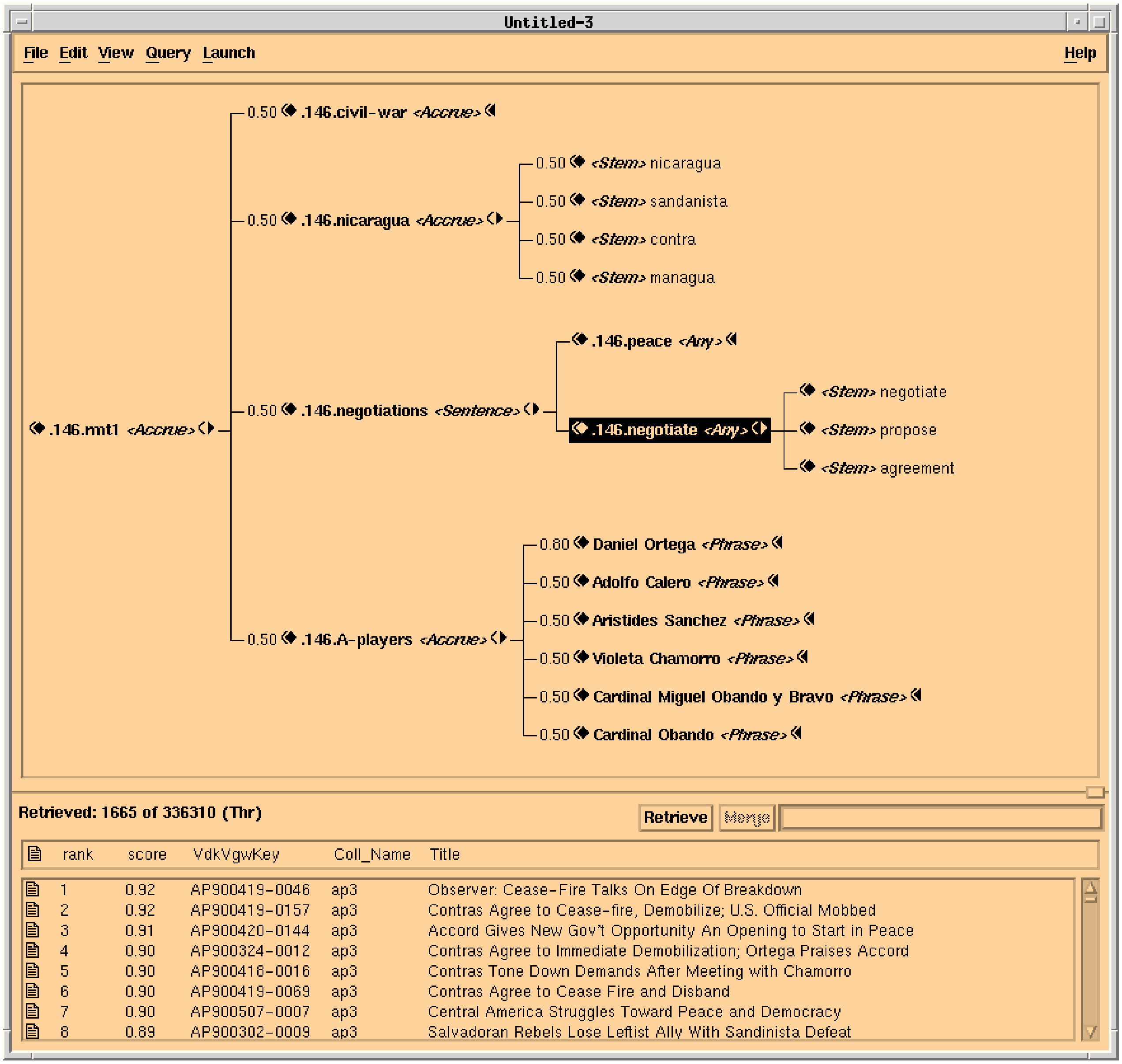

Figure 3: Typical TOPIC Interface Screen

The pull-down menus at the top of the window provide access to a large number of support functions, including the usual array of system level commands (the File menu), commands for editing and manipulating queries and associated data structures (the Edit menu), commands that control the way in which the query screen appears (the View menu), commands for setting parameters associated with the query itself (the Query menu), and commands that are calls to applications outside of TOPIC proper (the Launch menu).

The query pane itself is interactive. All commonly required editing functions are available via a mixture of point-and-click, pop-up menus and drag-and-drop operations. Thus a searcher can quickly build a sophisticated query that can be viewed graphically. The interface also supports some other views of the query, including a, so called, simple query in which users can enter unrestricted query terms, and a form-based query in which constraints can be placed on the searches to be performed on the fielded components of the documents.[6]

The results list is also configurable, but is shown with the default for the TREC experiment. That is, we show the document rank, its relevance score, its ID, the collection it comes from, and its “title.” To view a document, the searcher simply double-clicks on required item and the corresponding full-text of the document is displayed, with highlights, in a separate window.

We collected little data on the use of specific features in TOPIC. As we would predict, the anecdotal evidence suggests that the experienced TOPIC users made use of more features than the novices, but we can draw no specific conclusions from the limited data we have.

A.2 The Focus Topic

An important component of the interactive test for TREC-3 was the idea of using topic 122 as a "focus topic." For this focus topic we provide some specific details about query construction behavior, as well as a closer examination of the actual query developed.

Unfortunately, a logistical error in collecting the queries resulted in an intermediate query being used for the generation of the official TREC-3 results (i.e., the TOPIC2 results set), and so we present here an analysis of the final query the searcher actually produced. Therefore, of course, the performance measures differ from those officially reported.

The statement of information need for topic 122 is:

```
<dom> Domain: Medical & Biological
<title> Topic: RDT&E of New Cancer
  Fighting Drugs
<desc> Description:
Document will report on the research,
  development, testing, and evaluation
  (RDT&E) of a new anti-cancer drug
  developed anywhere in the world.
<narr> Narrative:
A relevant document will report on any
  phase in the worldwide process of
  bringing new cancer fighting drugs to
  market, from conceptualization to
  government marketing approval.  The
  laboratory or company responsible for
  the drug project, the specific type of
  cancer(s) which the drug is designed to
  counter, and the chemical/medical
  properties of the drug must be
  identified.
<con> Concept(s):
1. cancer, leukemia
2. drug, chemotherapy
```

We see that this is a complex topic that requires a document to contain a number of specific data items if it is to be considered relevant. Most information retrieval systems can approach this topic only indirectly, and TOPIC is no exception.

The initial formulation of the query (i.e., the one used to generate the results in the TOPIC1 results set) is essentially disjunctive with two major components — one that searches for variations of the "new anti-cancer drug" concept; and one that searches for information related to anti-cancer drugs, including the names of specific known drugs, the names of various cancers, the names of agencies concerned with regulating drugs, and synonyms for "treatment" and "research."

The final formulation of the query retains the basic elements of the initial query, but reorganizes them so that a retrieved document must contain evidence for the "new anti-cancer drug" concept (i.e., a necessary condition), and thereafter the degree of relevance is a function of the amount of related information found. This is modelled in TOPIC using a proprietary construct that can be thought of as a weighted conjunct.

The performance of these two formulations on the training data (i.e., the performance the searcher was responding to in their interaction with the system) are shown below. The scores for the initial formulation are labelled TOPIC0; the scores for the final formulation are labelled TOPIC00.

```
Queryid (Num):       122
                   TOPIC0      TOPIC00

Total number of documents over all queries
    Retrieved:     1000        1000
    Relevant:       119         119
    Rel_ret:         71          72
Interpolated Recall - Precision Averages:
    at 0.00       1.0000      1.0000
    at 0.10       0.4828      0.5517
    at 0.20       0.3429      0.3871
    at 0.30       0.2628      0.2975
    at 0.40       0.1582      0.2134
    at 0.50       0.1202      0.1760
    at 0.60       0.0000      0.0940
    at 0.70       0.0000      0.0000
    at 0.80       0.0000      0.0000
    at 0.90       0.0000      0.0000
    at 1.00       0.0000      0.0000

Average precision (non-interpolated) over
  all rel docs
```

6. The simple query mode is mostly used in the initial stages of query development when the searcher is more concerned with determining an appropriate set of retrieval terms than with the structuring of these terms. Once a reasonable set of terms have been identified, searchers usually switch to the graphical mode. The form query was not especially useful for TREC-3, since we provided only a limited number of searchable fields (i.e., the Coll_Name and Title fields). Indeed, an analysis of the TREC-3 data shows that no searcher used them in the final formulations of the submitted queries.

```
                    0.1782      0.2226

  Precision:

 At    5 docs:   0.6000      0.8000
 At   10 docs:   0.6000      0.8000
 At   15 docs:   0.6000      0.6667
 At   20 docs:   0.5000      0.5500
 At   30 docs:   0.4667      0.5333
 At  100 docs:   0.3000      0.3200
 At  200 docs:   0.2050      0.2300
 At  500 docs:   0.1200      0.1280
 At 1000 docs:   0.0710      0.0720

R-Precision (precision after R (= num_rel
 for a query) docs retrieved):
   Exact:        0.2857      0.2941
```

Inspection of these data shows that the revised query did indeed outperform the initial query, although not dramatically so. Overall recall remained the same, but average precision and precision@N were significantly better.

The searcher reports that this was a difficult query to develop and, although many variations were tried,[7] there was no query that generated a significantly higher recall while at the same time preserving "reasonable" precision. The searcher felt that the precision improvements exhibited by the selected variant were significant enough that the lack of recall improvement was "acceptable."

The main difficulty with this topic is, of course, that it requires relevant documents to describe a process (i.e., the development and marketing of a new drug) and identify specific entities involved in that process (i.e., the drug company, the properties of the drug, and its target). Since the process is not completely described by the information need statement (and was not known to the searcher) and there is no enumeration of the companies, drugs and cancers, the searcher developed a query that attempts to model various kinds of collateral indicators. As we would expect, this was a difficult, and essentially non convergent, task that required the searcher to amass a collection of relatively weak indicators of relevance (i.e., search terms). Given that no one group of indicators dominates, finding the right combinations and weights reduces to a search in a high-dimensional space — something humans are not very good at. This is, we believe, the reason why the searcher found this to be a hard topic, and ultimately found it difficult to build a satisfactory query.

7. The searcher's notes show that twelve major variants of the query were tested. A variant could include either a restructuring or a redefinition of base-level query terms, or both, and was marked as such if the searcher preformed a complete retrieval (i.e., ran the query over the entire training corpus) and then generated set of retrieval results using the TREC evaluation program.

The performance of these two queries on the actual routing experiment are shown below. The original query generated the results labelled TOPIC1, and the final query generated the results labelled TOPIC2.

```
Queryid (Num):       122
                   TOPIC1      TOPIC2

Total number of documents over all queries
    Retrieved:     1000        1000
    Relevant:        63          63
    Rel_ret:         61          44
Interpolated Recall - Precision Averages:
    at 0.00        1.0000      1.0000
    at 0.10        0.9091      0.9231
    at 0.20        0.7000      0.8235
    at 0.30        0.2159      0.3231
    at 0.40        0.1561      0.2593
    at 0.50        0.0862      0.0831
    at 0.60        0.0862      0.0588
    at 0.70        0.0862      0.0000
    at 0.80        0.0759      0.0000
    at 0.90        0.0715      0.0000
    at 1.00        0.0000      0.0000

Average precision (non-interpolated) over
 all rel docs
                   0.2900      0.3026

  Precision:

 At    5 docs:   0.8000      0.8000
 At   10 docs:   0.9000      0.9000
 At   15 docs:   0.6667      0.8000
 At   20 docs:   0.7000      0.7000
 At   30 docs:   0.5333      0.5333
 At  100 docs:   0.1900      0.2500
 At  200 docs:   0.1400      0.1400
 At  500 docs:   0.0800      0.0700
 At 1000 docs:   0.0610      0.0440

R-Precision (precision after R (= num_rel
 for a query) docs retrieved):
   Exact:        0.2857      0.3175
```

What is striking about these results is the very significant reduction in the number of relevant documents retrieved by the revised query. In contrast, the precision measures of interest (i.e., average precision, precision@10, precision@30, and R-precision) are either the same or higher for the revised query.

These data suggest, perhaps, that the reformulated query was a good model of those documents in the test collection that were similar in character to the training data, but that the necessary condition built into the revised query was too restrictive. However, as of this writing, we have not done a detailed enough analysis of the results to determine the exact causes for the low recall score.