Rutgers' HARD and Web Interactive Track Experiences at TREC 2003

N.J. Belkin, D. Kelly, H.-J. Lee, Y.-L. Li, G. Muresan, M.-C. Tang, X.-J. Yuan, X.-M. Zhang
School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ 08901
[belkin, diane, hyukjinl, lynnlee, muresan, muhchyun, xjyuan, xzhang]@scils.rutgers.edu

1 Introduction

This year, members of our group, the Information Interaction Laboratory at Rutgers, SCILS, participated in the HARD track, and in the Interactive Sub-track of the Web track. Since there were no points of commonality between the two separate investigations, we describe and present the results and conclusions for each separately.

2 The HARD Track

2.1 Introduction and hypotheses

The goal of our work in the HARD track was to test techniques for using knowledge about various aspects of the information seeker's context to improve IR system performance. We were particularly concerned with such knowledge which could be gained through implicit sources of evidence, rather than explicit questioning of the information seeker. We therefore did not submit any clarification form[1], preferring to rely on the categories of supplied metadata concerning the user which we believed could, at least in principle, be inferred from user behavior, either in the past or during the current information seeking episode. To this end, based on the training data supplied and our previous research, we attempted to test the following hypotheses:

H1: People who are familiar with a topic will want to see documents which are detailed and terminologically specific; people who are unfamiliar with a topic will want to see general and relatively simple documents. This we operationalized by promoting the value of documents which scored toward the unreadable end of readability scales for people highly familiar with the topic, and by promoting the value of documents which scored toward the easily readable end of the scales for people unfamiliar with the topic.

H2: Different document genres can be identified by their vocabul**aries**. This we operationalized by constructing language models for all the retrieved documents for each training topic and for just the completely relevant documents for each topic. We then identified words which occurred with greater than expected probability, based on the entire topic language model, in the relevant documents, for all topics which had the same genre. These words were considered to be indicators of the genre. We added the words associated with a particular genre to queries for topics which requested that genre.

H3: Certain document sources will be relevant, or not, to different desired genres. This we operationalized by promoting documents from certain sources to the top of the retrieved list for topics with some genres, by removing documents from some sources entirely from the retrieved list for topics with some genres, and by demoting the value of documents from some sources in the retrieved list for topics with some genres.

H4: If there are texts which the information searcher has identified as relevant to the topic, using them as the basis for automatic query expansion will improve retrieval performance. This was operationalized by choosing terms for query expansion from the relevant texts, based on a combined ranking formula.

H5: If the desired granularity of the retrieval result is passage, then the retrieved documents should be ranked on the basis of their best passage, rather than on the document as a whole. This was operationalized by using the InQuery best passage ranking function.

Our official submission was with queries constructed on the basis of hypotheses 2, 4 and 5.

Our basic IR system was InQuery, version 3.2, obtained from the Center for Intelligent Information Retrieval, University of Massachusetts (http://ciir.cs.umass.edu) using its default indexing, query processing and retrieval algorithms. The queries for our baseline run were constructed using both title and description fields from the topics, and were just the weighted sum of the stemmed, non-stoplist words from the title and description fields. These queries were then used as the basis for our experimental runs, with them, or their results, modified according to the metadata, as described in section 2.2, below.

2.2 How metadata about the searcher was used

The experimental condition of the HARD track was for each site to submit at least one baseline run for the set of 50 (eventually 48) topics, using only the title and (optionally) description fields for query construction. The results of the

[1] See Allan, this volume, for detailed information about the goals and conditions of the HARD track.

baseline run(s) were compared with the results from one or more experimental runs, which made use of the searcher metadata that was supplied, and of a clarification form submitted to the searcher, asking for whatever information each site thought would be useful in improving search results. We used only the supplied metadata, for the reasons stated in section 2.1, and especially because we were interested in how to make initial queries better, rather than in how to conduct a dialogue with a searcher. There were five categories of searcher metadata for each topic (not all topics had values for all five): Purpose, Genre, Familiarity, Granularity and Related text(s), which were intended to represent aspects of the searcher's context which might be useful in tailoring retrieval to the individual, and the individual situation. We made the assumption that at least some of these categories would be available to the IR system prior to (or in conjunction with) the specific search session, either through explicit or implicit evidence. Therefore, for us the HARD track experimental condition was designed to test whether knowledge of these contextual characteristics, and our specific ways of using that knowledge, would result in better retrieval performance than a good IR system without such knowledge.

We understood that there would be, in general, two ways in which to take account of the metadata. One would be to modify the initial query from the (presumed) searcher, before submitting it for search; the other would be to search with the initial query, and then to modify (i.e. re-rank) the results before showing them to the searcher. We used both of these techniques in taking account of the different types of metadata.

Knowledge of the purpose of a search (i.e. the searcher's general goal) has long been understood to be important for human search intermediaries in tailoring a search to the specific user (cf. Belkin, 1984). Whether such knowledge could be used effectively in a direct end-user IR system is still an open question. Unfortunately, we were unable to investigate this issue in this experiment. One reason for this is that the training data that were supplied in the HARD track did not have sufficient variety on this characteristic for us to investigate different hypotheses about how to take account of it; another is that the types of purpose that were identified did not immediately suggest how they could be used.

Desired genre for the results of a search has also been identified as potentially significant in improving search performance (e.g. Rauber & Müller-Kögler, 2001). In this case, we had two hypotheses. One was general: that the genre of a document could be identified by its vocabulary. This hypothesis we operationalized in the following way. For the training data, we constructed a language model[2] based on the top 100 documents retrieved by our basic query for each topic, and a language model based on all of the documents which were evaluated as both topically relevant, and satisfying all of the metadata conditions with respect to that topic. We then identified those words which appeared with a significantly higher probability in relevant documents than in all retrieved documents, for each topic associated with each specific genre. We also identified those words which were significant in the relevant documents, but had a low probability of being generated by the language model of the retrieved documents. Using these two lists, and given the nature of the metadata, we were able to identify some words which seemed to be indicative of the genre class, Overview. These words: *one, two, three, year, last, more, total, average, historically, spanning, surveyed, trends*; were added to the baseline queries for all topics which specified Genre as Overview, using the InQuery "or" operator.

The second hypothesis for genre was based on specifics of the HARD collection. The HARD database consists of the AP Newswire, the New York Times, the Xinghua newspaper (in translation), the Federal Register and the Congressional Record. We noted that documents satisfying the Genre category of Administrative were almost certainly to be found in the Federal Register or the Congressional Record. For such topics, we therefore submitted the basic query, and increased the value on which the document rank was based (the Retrieval Status Value – RSV) for all Congressional Record and Federal Register documents as follows:

$$new\ RSV = 1 + original\ RSV \quad (1)$$

This had the effect of placing all CR and FR documents at the top of the retrieved list, in their original order with respect to one another. We also noted that the Genre category Reaction would almost certainly never be satisfied by a document from the Federal Register collection, and was most likely to be satisfied by documents from news databases. We therefore deleted all Federal Register documents from the results lists for topics with Genre = Reaction, and demoted the value of Congressional Record documents according to the following formula:

$$new\ RSV = original\ RSV - 0.5(original\ RSV) \quad (2)$$

Familiarity with a topic has been identified as having a significant impact on relevance assessments and on how interactive IR searches are conducted (e.g. Kelly & Cool, 2002), and it is easy to imagine various ways in which familiarity would impact understanding and usefulness of a document to a person. We hypothesized that people familiar with a topic would not only be able to read and understand technical and detailed documents on the topic, but that they also would prefer those to more general documents on the topic. On the other hand, people who are unfamiliar

[2] Using the language modeling toolkit at http://mi.eng.cam.ac.uk/~prc14/toolkit.html

with a topic might prefer more general documents, and might not be able to comprehend technical ones. Failing any better ideas, we decided to use readability as a measure of technicality/generality; the less readable, the more technical, the more readable, the more general. Although there was insufficient variety on this characteristic in the training data for our hypothesis to be tested on it, we did compute the readability of a systematic random sample of the HARD collection. This led us to an additional hypothesis: that some documents are too simple to read or too unreadable to be of use to anyone searching in this collection. We therefore implemented the following procedure for taking account of familiarity.

The readability of each of the top 1200 documents retrieved by a query to the collection was computed, using three widely used measures. The measures were Fog index, Flesch reading ease score, and Flesch-Kincaid grade level score, computed using algorithms implemented in the PERL programming language by Kim Ryan in 2000[3]. All documents which had all three readability scores at or below, or at or above extreme outlier values for the collection as a whole (as estimated by our sample of the collection) were discarded from the results. Then, for all topics which had a readability level of 4 (meaning very familiar), the RSV was increased for documents which had a readability score greater than (meaning less readable) or equal to 3 standard deviations above the mean as follows:

$$new\ RSV = original\ RSV + 0.2(original\ RSV) \quad (3)$$

For all topics which had a familiarity level of 1 (meaning no familiarity), the RSV for documents which had a readability score less than (meaning very readable) or equal to 3 standard deviations below the mean were promoted according to equation (3).

Granularity of response was a category of metadata to which we paid relatively little attention, primarily because we did not have the capability for effective passage and sentence-level retrieval. However, we made the assumptions that documents with highly relevant passages might have those passages near the beginning of the document, or that such passages would be easy to spot in the document. Then we addressed the Granularity category of Passage by submitting the queries for all such topics using InQuery's passage-level ranking of retrieval results rather than whole-document-based ranking, with a passage length of 200 words, approximating a paragraph.

Finally, we used the Related Text metadata as the basis for query expansion (QE) of the baseline queries for all topics which specified related texts. We did not use these texts for query term re-weighting, and we simply added the QE terms to the basic weighted sum query. The terms added to a query were determined by using three different QE term-ranking measures on the set of relevant texts, combining the rankings according to the median rank, and then selecting the top 10. We decided on this method based on results reported by Carpineto, Romano & Giannini (2002), which suggest that using different QE ranking techniques and then combining them leads to better retrieval performance than using any single QE ranking technique. We ranked according to the following three formulae:

$rank = t$, *with ties being resolved according to DF, lowest DF value first;*

$rank = [(t/R) - (t/DF)] / t/DF;$

$rank = (t/R) \times log[(t/R)/(t/DF)]$;

in which t represents number of occurrences of the term in the relevant documents; R represents total number of term tokens (i.e. the number of *different* words) in the relevant documents; and DF represents total number of documents in the collection with the term..

We planned to apply the different techniques for taking account of the various metadata types in sequence, combining them all into one single query modification plus results re-ranking as follows:

Baseline query + relevant text QE + Overview words + passage-level ranking = results list 1

Results list 1 + Administrative re-ranking + Reaction re-ranking + Familiarity re-ranking = final result list

Unfortunately, for a variety of reasons, we were able to complete this process only as far as results list 1 in time for the official submission. This is the basis for the results reported below.

2.3 HARD results

Our baseline results were rather good, and substantially above the median of the experimental results for all systems. This is likely to be a result of our using both title and description for our queries; it seems likely that most other sites used title only, or title plus some form of pseudo-relevance feedback or other query expansion technique. Of more interest, of course, are our experimental results.

---

[3] Available online: http://aspn.activestate.com/ASPN/CodeDoc/Lingua-EN-Fathom/Fathom.html#SYNOPSIS

With respect to experimental results from all sites participating in the HARD track, Rutgers did quite well. Figure 1 indicates, for each topic, the amount above or below the median value of the Rutgers results for both R-precision and average precision.

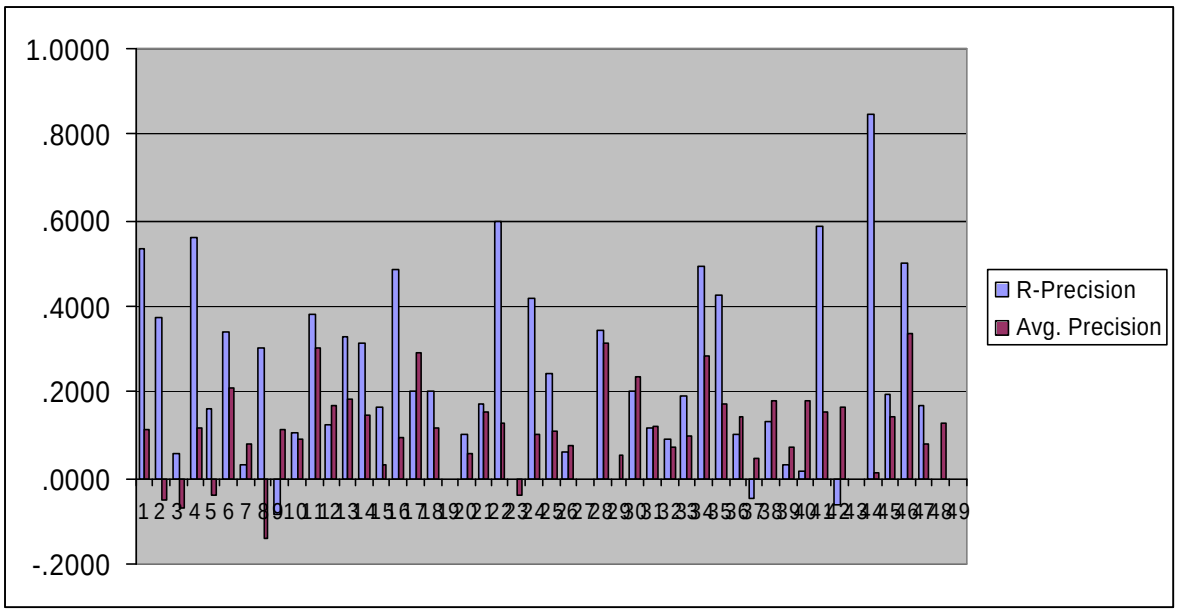

Figure 1. Difference between median values and Rutgers results for R-precision and Average Precision

Table 1 shows roughly the same data, indicating how many times the Rutgers results were best, above the median, at the median (M), and below the median for three performance measures (Rutgers was not worst for any topics).

Measure	Best	Above M	At M	Below M
Rel. Ret. @ 10	10	23	13	2
R-precision	4	32	8	4
Average Precision	3	39	3	3

Table 1. Rutgers' results compared to all results for experimental run.

Unfortunately, this comparison to everyone else does not really tell the full story. In fact, since the goal of the HARD track is to use metadata to improve over the baseline, it is much more important to look at that comparison. Here, things do not look so good. In fact, as table 2 indicates, performance on almost all measures was slightly lower for our experimental run (called Rutmeta) than for our baseline run (called rutbase2), when summarizing over all topics. Although the differences are clearly not significant, they are somewhat disheartening.

Run	Precision @ 10	R-precision	Avg. Precision	Rel. Ret.
rutbase2	0.4750	0.3451	0.3186	3736
Rutmeta	0.4750	0.3308	0.3019	3728

Table 2. Mean values of performance measures for baseline and experimental Rutgers runs.

Fortunately, this again does not tell the whole tale. If the results are compared on a topic-by-topic basis, and cumulated as in table 3, then we see that for three out of the four measures, the baseline did better than the experimental run a few times, but for average precision, the experimental run did better on 26 out of the 48 topics, and was equal for three.[4]

[4] For four topics, there was no metadata used at all, so these are not counted.

	Rel. Ret. @ 10		R-Precision		Avg. Precision		Rel. Ret	
	Rutmeta	rutbase2	Rutmeta	rutbase2	Rutmeta	rutbase2	Rutmeta	rutbase2
Better	11	15	16	19	26	17	12	17

Table 3. Topic-by-topic comparison of performance between baseline and experimental runs.

Although we do not have results which can conclusively indicate what effect each of our different techniques had on performance, we can look at some aspects of this issue. The various topics had different combinations of metadata that we used in our official experimental run, so that there are instances of each technique used separately, and the techniques used in various combinations. Table 4 indicates the effect of the different techniques by displaying for how many of the four evaluation measures using the particular metadata technique, or combination of metadata techniques, the technique did better, the same as, or worse than just the baseline. Entries in the 3 leftmost data columns indicate some advantage to having used the metadata; the fourth and fifth data columns indicate no real difference between metadata and baseline, and the sixth and seventh data columns indicate a distinct disadvantage to using the metadata. These data suggest that there was some overall advantage to enhancing the baseline query by using relevant text query expansion in combination with overview query expansion, and, if we disregard the "no difference" values, that using metadata had an overall advantage of better performance on 19 topics compared to 15 topics with better performance in the baseline condition. We still need to figure out how it happened that a topic to whose query we thought we had done nothing turned out to perform worse in the experimental condition than in the baseline.

Metadata	3 or 4 >	2 > 2 =	2 > 1 = 1 <	3 or 4 =	2 > 2 < 2 < 2 =	3 or 4 <	Total
None (3)						1!	4
QE only	4	1	1	2	4	6	18
Passage only	1						1
Overview only				2		1	3
QE + P			1	1	1	3	6
QE + O	5		3			1	9
P + O				1	1		2
QE + P + O	1	1	1			2	5
TOTALS	11	2	6	6	5 1	14	48

Each column indicates the number of topics for which using metadata resulted in the specified number of evaluation measures being better, equal to, or worse than the baseline .
> means Meta better than baseline; = means Meta same as baseline < means Meta worse than baseline
QE is query expansion; Passage (P) is ranking by best passage; Overview (O) is adding "overview vocabulary" to queries.

Table 4. Effect of application of different metadata information to baseline queries.

We also did several runs in which we tested the effect of applying only one category of metadata at a time to the baseline run. The results are displayed in Table 5, where it is easy to see that using Overview query expansion and our version of Passage retrieval had no effect. However, both Genre using the source, and Query expansion had positive effects on performance. Although the differences in performance levels are typically not great for these two, the number of topics positively affected by these two treatments was substantially greater on several of the measures.

2.4 Discussion and conclusions on the HARD results

Although the average performance of our official run using metadata is somewhat lower than our baseline run, more detailed analysis suggests that we did indeed gain some advantage from using the metadata to modify the baseline queries, in some respects. In particular, performance as measured by average precision was improved for well over half the topics, and there appears to be some advantage to the relevance feedback-like query expansion techniques. The language model-based genre technique did not work well, however. Of course, the ways in which we used the metadata to modify rankings and queries were quite ad hoc, and without real theoretical justification, which could go some way toward explaining negative results. We are still not in a position to evaluate properly the effects of each of the techniques which we have proposed on retrieval performance, nor of their complete combination, nor are we able to respond with any level of confidence to our initial hypotheses. We intend to perform further studies in which we compare all of the different techniques, and vary their parameters, in order to address this problem.

Metadata	Overview		Genre (Passage)		Query Expansion		Genre (Source)	
Number of topics	20		14		38		12	
Condition	base	meta	base	meta	base	meta	base	meta
Rel. Ret.	3736	3732	3736	3645	3736	3715	1062	1046
Number better*	1	1	3	7	15	11	3	4
Avg. Prec.	0.3186	0.3196	0.3186	0.3041	0.3186	0.3187	0.2538	0.2666
Number better	3	1	9	4	11	22	2	4
Prec. @ 10	0.4750	0.4667	0.4750	0.4646	0.4750	0.4938	0.4250	0.4917
Number better	2	0	5	5	11	11	0	3
R-Prec.	0.3451	0.3458	0.3451	0.3284	0.3451	0.3475	0.2945	0.3178
Number better	1	1	9	2	11	17	2	3

*Number of topics for which the condition had better results. When the two add to less than the total topics, all others were equal.

Table 5. Performance of single metadata treatments compared to baseline.

3 The Web Interactive Track

3.1 Introduction and hypotheses

This year the interactive TREC experiment was set up as part of the Web track and was built around the topic distillation task: finding a list of key resources for a particular topic, concentrating solely on websites as resources[5]. In the interactive sub-track, the searchers' task was to construct such a resource list for each of a set of broad topics, through interaction with an information access system[6]. The purpose of the experiment was to investigate whether the human capacity to interpret and summarize can beat machine algorithms at the topic distillation task. Apart from the direct comparisons of results, the observation of the human searchers' behavior could potentially offer clues to improving topic distillation algorithms.

We investigated the role that **the layout of search results** plays in supporting human searchers executing topic distillation tasks. Success was measured in terms of accuracy and precision, operationalized as coverage and overlap, so the searcher was expected to find documents that provide information on as many distinct aspects of the assigned topic as possible, with as little overlap between them as possible. Our hypothesis was that using the structure of the domain and of the document corpus in order to organize the search output, would help identify aspects of the search topic in different sub-domains of the document collection, would reduce the searchers' cognitive load and would produce better results than the classic hit list. We tested this hypothesis by using two user interfaces for the Panoptic search engine, one with a simple list output, and the second with documents clustered based on common URL elements.

The experimental (or hierarchic) interface, depicted in Figure 2 and described in Box 1, grouped the search results based on commonality of URL parts (sub-domain and path) and displayed them in a one level tree. The groups of hits were ranked based on the Panoptic rank of their top document; the Panoptic ranks were also used to sort hits within each group. The structured layout determined us to take two design decisions that go against common Web search engine result arrangements. Firstly, we reckoned that "More results" or "Next page" would be either ambiguous or confusing, so we did not provide such functionality. Instead, the sets of search results contained 30 hits, which was considered sufficient for the topic distillation task: if no relevant document can be found in the top 30 hits, then a query formulation is probably more appropriate than a request for more hits. Secondly, also in order to avoid confusion, the actual ranks were not displayed in the hierarchic output, but the subjects were explained the ranking scheme.

The baseline (or linear) interface was almost identical, the only difference being the layout of the 30 hits: they were displayed in a list, with the ranking provided by Panoptic. For consistency, the ranks were not displayed, but the subjects were told that documents at the top of the list were more likely to be relevant.

We used the neutral version of Panoptic, so that the subjects' task would not be supported by a topic distillation algorithm; judging the relevance of retrieved documents and the completion of the topic distillation task was entirely based on the subjects' effort.

Apart from the measures of coverage and overlap, provided by NIST based on the assessors' relevance judgments, we planned to use a set of objective measures that indicate search effort such as time required to complete the task, number

[5] http://es.cmis.csiro.au/TRECWeb/guidelines_2003.html
[6] http://www.ted.cmis.csiro.au/TRECInt/guidelines.html

of iterations (or queries submitted), number of documents seen[7], selected[8] and viewed[9], number of documents saved during the interaction and number of documents kept[10]. We also prepared questionnaires in order to measure subjective measures of success such as user satisfaction and perception of success, and to investigate the correlation between success and measures such as familiarity or expertise with the topic, search expertise etc.

We were also interested in continuing previous years' investigation by looking at the effect that the query formulation panel and the instructions provided to the subject have on the syntax, length and specificity of the queries submitted. As time and resource constraints did not allow us to build another two user interfaces and run more subjects, we have no rigorously tested results. However, observations of the subjects and comparisons to last year's experiment allowed us to draw some anecdotal conclusions.

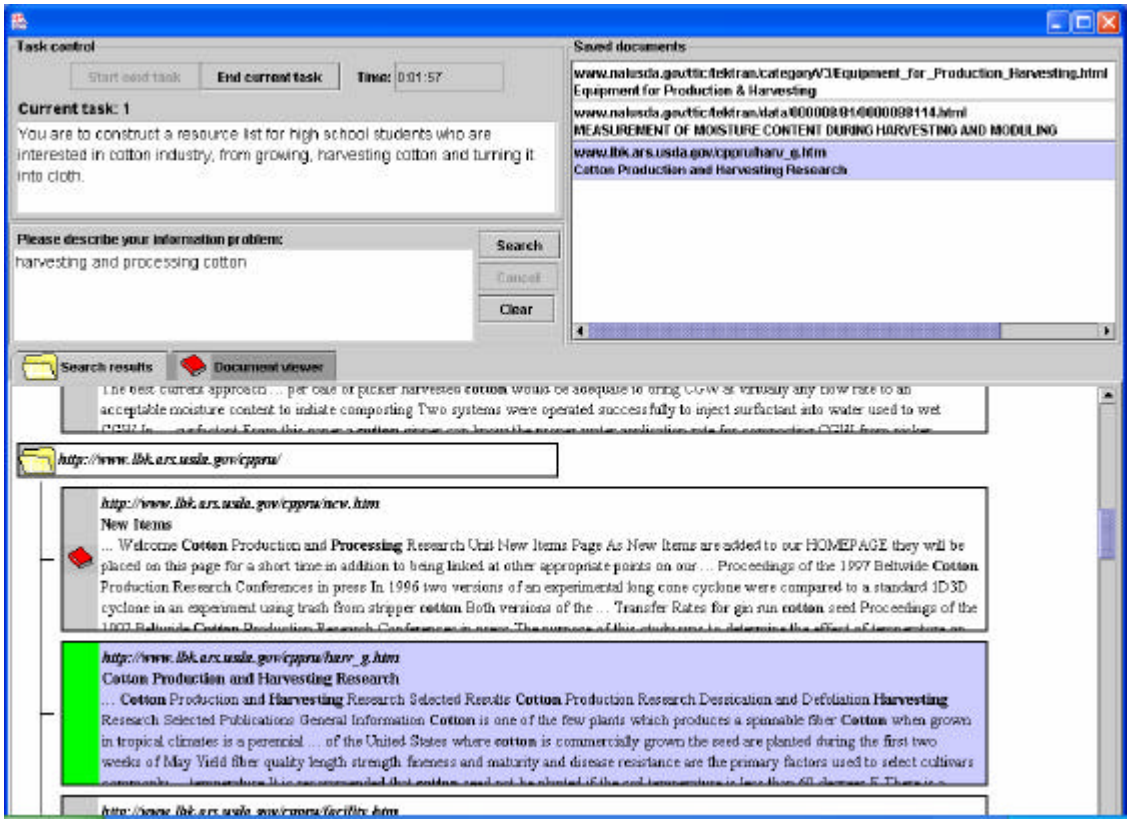

Figure 2: The experimental user interface.

[7] Document surrogates seen while scrolling through the search results.
[8] Documents selected from the set of search results output by Panoptic. These are a subset of the set of documents seen.
[9] Documents either selected from the hit set, or obtained by following links in the Document Viewer, or by editing the URL and loading the specified webpage, if available in the .gov collection.
[10] Saved documents could be unsaved if the subject found better documents.

Box 1. Description of the user interfaces:

The **Task Panel** allows the subject to start a task (which opens a log file), displays the text of the task, including the topic, keeps track of the time, and allows the subject to end the task (which closes the log file).
The **Query Panel** encourages the subject to describe the information problem and provides sufficient space for several sentences.
The **Search Results Panel** displays the output from the Panoptic search engine, each hit being represented by a URL, a document title, and a summary. The subject can scroll and select documents for viewing in the Document Viewer. For improved usability, color coding is used to mark the currently selected document (visible by switching to the Document Viewer by clicking on the appropriate tab), the already saved documents and the already viewed documents.
The **Document Viewer** displays the full text of the selected document, allows the user to follow hyperlinks, to specify a URL and to request the loading of the specified document; it also allows the user to save the current document or to go back to a previously displayed document.
The **Saved Documents Panel** displays the URL and title of the saved documents. If the user clicks on one of the items in the panel, the corresponding document is displayed in a new window, above the Document Viewer, for comparison with the current document, so that the user can decide if there is overlap between the documents and which document is better. A saved documents can be unsaved if the user finds a better one as replacement, or reviews its relevance in view of the information retrieved.

3.2 The Interactive experiment

We had 16 subjects, volunteers mostly recruited from among Library and Information Science students. Eight were female and four male, and the ages were evenly distributed in the range 18-47. They all displayed a high level of experience with computers (6.44, 0.96), with WWW browsers (6.31, 1.01), with search engines (6.25, 0.93), and displayed a high level of confidence in being able to find information (6.00, 1.10)[11]. While this gave us confidence that the subject would easily learn and adapt to our user interfaces, it also made impossible any comparison between people with different levels of expertise. The experimental design was established by NIST, so the reader is referred to the relevant webpage[12], which also details the topics. Each subject conducted eight searches, four on the baseline and four on the experimental system. The order of systems and topics was rotated as described in the experimental design to minimize the effect of learning and tiredness on the result.

3.3 Data analysis and results

3.3.1 Objective measures

Each set of documents saved by each subject, while searching on each of the eight topics, was judged by two NIST assessors and given two scores by each: coverage (of the different aspects of the topic), ranging from 1 (very good) to 5 (very bad) and overlap (between saved documents), ranging from 1 (none) to 5 (way too much). Although there were significant differences in the reviewers' judgments, the conclusions drawn from comparing the linear and the hierarchic system were consistent. Even though a t-test failed to find a statistically significant difference, the data summarized in Table 6 indicates a tendency of linear display to be more conducive to better coverage and of hierarchal display to be more conducive to less overlap.

		Linear	**Hierarchy**
Reviewer 1	Coverage	2.67(1.72)	2.92(1.64)
	Overlap	2.59(1.23)	2.34(.96)
Reviewer 2	Coverage	2.58(1.73)	2.69(1.77)
	Overlap	2.44(.99)	2.25(1.05)

Table 6: Search results judged by expert reviewers

[11] The values in parentheses represent mean values and standard deviation on a 7-point Likert scale.
[12] http://www.ted.cmis.csiro.au/TRECInt/guidelines.html

A possible explanation of this result is that users of the baseline interface have no structure to support their exploration of the search results and therefore have to scan a larger number of documents to be satisfied with what they find. While more time-consuming and more cognitively demanding, this process has the potential to give a better coverage of a topic. On the other hand, the users of the hierarchic system have the option to direct their browsing at different sub-domains of the collection; once the user gets familiar with this kind of output, it is expected that the user would do more analysis, deciding what areas to explore, and less browsing, the result being less "direct interaction" and less overlap between content of saved documents.

Scanning all the documents in a collection has the potential for complete coverage of a topic, but is obviously not feasible; recall needs to be balanced by precision or effort. The slight increase in coverage shown by the linear system needs to be considered in the context of effort, measured in terms of time taken to search, number of iterations, and number of documents seen, selected and viewed. These measures are compared in Table 7.

Interaction Measures	**Linear**	**Hierarchy**
Iterations	4.19(2.85)	3.61(1.96)
Time (seconds)	618.53 (204.85)	575.80(117.81)
Number seen	42.78(22.12)	41.91(21.95)
Number viewed	11.81(4.52)	11.50 (5.02)
Number selected	10.64 (4.11)	10.25(4.35)
Number of ever saved	6.14 (3.08)	5.78(2.98)
Number of final saved	6.00 (3.00)	5.63(2.97)
Ratio of viewed to seen	.312 (.18)	.306(.19)
Ratio of selected to seen	10.64 (4.11)	10.25 (4.35)

Table 7: Interaction measures by display modes

Even if the difference is not statistically significant, the data in this table indicates a tendency that appears to confirm our hypothesis and expectations: the hierarchic system is conducive to less interaction.

Based on previous years' experiments, which indicated a negative correlation between user satisfaction with a system and the amount of interaction (Belkin et al, 2003a), the results from our objective measures would predict that the users would prefer the hierarchic system. Other experimental results have indicated that users like to have control over the interaction (Koenemann & Belkin, 1996); this provides another reason for us to expect the hierarchic system to be favored by users, as it allows the searcher more navigational control. Let us see if our subjective measures confirm our expectations.

3.3.2 Subjective measures

3.3.2.1 Direct comparison

The exit questionnaires provide a direct comparison between the two systems: the subjects were asked which system they found easier to learn, easier to use, which system they felt supported the task better, and which system they liked more overall. The results are as shown in Table 8:

	Linear	**Hierarchical**	**No difference**
Easier to learn to use	4	1	11
Easier to use	3	8	5
Support your tasks better	3	9	4
Like the best overall	2	10	4

Table 8: Direct system comparison (frequencies)

The results show that most of the subjects (11) perceive no difference between the linear and the hierarchic system with respect to which one is easier to learn to use. On the other three questions most of the subjects (8, 9, 10 respectively) preferred the hierarchical system. A Chi-Square test indicates that the skewness of the distribution of subject perception is statistically significant in terms of ease to learn ($?^2$ (2, N=16) = 9.875, p<.01) and overall preference ($?^2$ (2, N=16) = 6.500, p<.05) and not quite significant in the other cases. We can conclude that the systems are perceived as similar in ease to learn and that people prefer the hierarchic output. Apart from the layout of the display, the two systems were identical, which explains that the subjects found no real difference in learning to use them and in using them. As expected, they clearly preferred the hierarchic display.

3.3.2.2 Indirect comparison

An indirect comparison between systems was provided by answers to questionnaires administered after a subject finished using a system. The questions focused on the searchers' perception of the system with regard to ease to learn, ease to use, understanding of how to use the system, and usefulness in helping accomplish the search tasks. The subjects answered by assigning scores on a 1 – 7 Likert scale, and these scores obtained by the systems were compared by a t-test. No statistical difference was observed overall ($t(30) = -.048$, $p>.05$), or in terms of ease to learn ($t(30) = -.425$, $p >.05$), ease to use ($t(30) = -.116$, $p>.05$), clarity of the conceptual model ($t(30) = .227$, $p>.05$) or usefulness for the search task ($t(30) = -.374$, $p>.05$).

Another indirect comparison between systems was provided by answers to questionnaires administered after each of the eight searches. The questions focused on the subjects' perception of the task completion and the quality level that was achieved on each task:

- "Do you think the resource list you just constructed focuses on the topic well?"
- "Do you think the resource list you just constructed provides a good coverage of the topic?"
- "Do you think the resource list you just constructed will be helpful for those people who are interested in this topic?"

	Linear	**Hierarchy**
Compiled list helpful to others	5.09(1.57)	4.88(1.59)
Compiled list focuses on the topic	4.95(1.58)	4.95(1.59)
Have enough time to do search	4.75(1.62)	4.38(2.01)

Table 9: Subjects' perception of search results

The results in Table 9 indicate that the sets of documents saved with the linear system tend to be slightly better, which correlates with the slightly better coverage observed in the objective measures. However, a t-test ($t(126) = .324$, $p>.05$) shows no significant.

Table 10 shows the correlations between a few task-related factors and the subjects' subjective search performance (as measured by a scale constructed from their responses to the post-search questions on the extent to which the compiled list is helpful to others, covers the topic, and focuses on the topic). Data on these factors were collected both before and after each search. The results show that all of them are highly correlated with the subjects' subjective search performance. This suggests that these factors may be critical to impact the subjects' search performance

		Correlation with subjective search performance
Before the search	Familiarity with the topic	.420**
	Expertise on the topic	.423**
	Perceived amount of available information on topic	.263**
After the search	How easy the task was perceived to be	.760**
	Enough time for this task	.660**

**Correlation is significant at the 0.01 level (2-tailed).

Table 10: Correlation of subjective assessment of search performance with task-related factors

3.4 Discussion of the interactive results

3.4.1 Query formulation

Last year we investigated two query-formulation modes. In the former, the experimenter and the text displayed in the user interface encouraged users to submit keywords. In the latter experimental mode, the subjects were specifically asked to use sentences to describe their information need and were provided sufficient space to do so. The experimenters' insistence and their demonstration of describing information problems in sentences, combined with the parenthetical statement "(the more you say, the better the results are likely to be)" had effect on the subjects' behavior: they did follow the instructions and did write sentences. These sentences proved to provide longer queries and fewer iterations, and to generate more satisfaction with the search outcome (Belkin et al., 2003b). This year, the Query Panel of our user interface was nearly identical to that from the second mode of last year and provided the same amount of space, suitable for writing several sentences. The difference was that the parenthetical statement was removed from the Query Panel, which was reduced to "Describe your information problem" and the subjects were not specifically asked

to write sentences. The result: no subjects generated any sentences. Very familiar with Web searching, the subjects seemed to enter "Google mode": they ignored the instruction from the screen and typed instead keywords, as they are used to. Consequently, the query length distribution (mean 3.04 (st.dev. 1.25) including stopwords and 2.72 (0.87) without stopwords) was surprisingly low compared to the expectations created by last year's experiment. Another explanation for this behavior may be related to the fact that, unlike last year, the topic descriptions were rather "naively" constructed, in order to be appropriate for the automatic tasks of the Web track. The essential topic keywords were present in the topic description, so most users copied and pasted the keywords into the query box, rather than having to generate them based on a problem and context description.

3.4.2 User comments

In the exit interviews, many subjects praised the capacity of the hierarchic organization to separate the different sub-domains of the collection and therefore different aspects of the topic at hand. The structured output saved them from having to mentally organize the hits and judge the overlap between their content; this was perceived as saving both time and cognitive effort. Such comments confirm our intuition that a structured display should support a structure-based task such as topic distillation.

Another feature mentioned often was the Saved Results panel, which helped users keep track of documents saved and allowed them to do side-by-side comparison between the currently examined document and already saved documents in order to compare their quality and the degree of content overlap.

Some comments indicated the need to improve the usability of the interfaces and the clarity of the underlying conceptual model. Despite the pre-experiment tutorial, some subjects did not notice the difference between the linear and the hierarchic display, as the indentation of the hierarchy was not seen as significant; others thought that the order of the hits in the display was random, rather than based on some probability of relevance score.

There were several complaints concerning the experiment settings: the constraint to limit the search to .gov documents invalidated many hyper-links, which frustrated most subjects; the time limit (10 minutes) put pressure on searchers and potentially generated un-natural behavior; thinking aloud impaired some users' ability to concentrate on the test.

3.5 Conclusions on the interactive experiment

Although it does not produce better coverage than the linear interface, the hierarchic interface seems to be conducive to less effort for the searcher: fewer iterations, shorter search sessions, fewer documents seen, selected and viewed. With regards to subjective measures, users perceived the hierarchic one as easier to use and better at supporting the topic distillation task. These results were not statistically significant. What was statistically significant is that the subjects perceived the two systems equally easy to learn and that they prefer the hierarchic display.

One advantage of the structured output, as suggested by the objective measures and highlighted by the users' comments, is the support for investigating different sub-domains of a document collection and consequently different aspects of a topic. The searcher does not need to make a cognitive effort to separate the search results into sub-domain, so the layout makes the interaction easier and more pleasant and more accurately supports the searcher's judgment on task completion.

This correlates with results obtained by CSIRO at TREC 2002: although the motivation to use a hierarchic organization was somewhat different, the structure imposed on the search output improved the retrieval performance in the case of complicated tasks, when relevant information needs to be gathered from various parts of the document collection (Craswell et al, 2002).

One direction in which we intend to continue our investigation is in displaying more than one levels of the hierarchic structure of webpages. While experiments with Cha-Cha[13] have shown promise, we are interested in whether combining navigation by browsing the hierarchic structure and following links to other parts of the hierarchy would help or confuse users.

4 Acknowledgements

Many thanks to Colleen Cool for her work on the HARD study, to Michael Coviello for his contributions to the interactive experiment, and to our volunteer subjects. The research reported here was supported in part by a Rutgers Research Council Grant to X.-M. Zhang, and by NSF grant No. 9911942. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

[13] http://cha-cha.berkeley.edu/

5 References

Allan, J (this volume) Overview of the TREC 2003 HARD track.

Belkin, N.J. (1984) Cognitive models and information transfer. *Social Science Information Studies*, vol. 4, nos. 2&3: 111-129.

Belkin, N.J., Cool, C., Kelly, D., Kim, G., Kim, J.-Y., Lee, H.-J., Muresan, G., Tang, M.-C., & Yuan, X.-J. (2003a). Interaction and query length in interactive retrieval. In D. Harman & E. Voorhees (Eds.), *The Eleventh Text Retrieval Conference (TREC2002)* (pp. 539-548). Washington, DC.: GPO.

Belkin, N.J., Cool, C., Kelly, D., Kim, G., Kim, J.-Y. , Lee, H.-J., Muresan, G., Tang, M.-C., Yuan, X.-J. (2003b) Query Length in Interactive Information Retrieval. In *Proceedings of SIGIR 2003* (pp.205-212). New York: ACM.

Carpineto, C., Romano, G. & Giannini, V. (2002) Improving retrieval feedback with multiple term-ranking function combination. *ACM Transactions on Information Systems*, v. 20 no. 3: 259-290.

Chen, M., Hearst, M., Hong, J., Lin, J. (1999) Cha-Cha: A System for Organizing Intranet Search Results, in Proceedings of the 2nd USENIX Symposium on Internet Technologies and SYSTEMS *(USITS)*, Boulder, CO, October 11-14, 1999.

Craswell, N., Hawking, D., Thom, J., Upstill, T., Wilkinson, R., Wu, M. (2002) TREC11 Web and Interactive Tracks at CSIRO. In E. Voorhees & D. Harman (Eds.) *Proceedings of TREC2002* (pp. 197-206). Washington, DC: GPO.

Kelly, D. & Cool, C. (2002) Effects of topic familiarity on information search behavior. In *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries – JCDL 2002* (pp. 74-75). New York: ACM.

J. Koenemann, J. & Belkin, N.J. (1996) A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In *Proceedings of CHI '96* (pp. 205-212) New-York: ACM..

Rauber, A. & Müller-Kögler, A. (2001) Integrating automatic genre analysis into digital libraries. In *Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL 2001* (pp. 1-10). New York: ACM.