WIDIT in TREC-2004 Genomics, HARD, Robust, and Web tracks

Kiduk Yang, Ning Yu, Adam Wead, Gavin La Rowe, Yu-Hsiu Li, Christopher Friend, Yoon Lee
School of Library and Information Science, Indiana University, Bloomington, Indiana 47405, U.S.A.
{kiyang, nyu, awead, glarowe, yuhli, cmfriend, yoonlee}@indiana.edu

1. Introduction

To facilitate understanding of information as well as its discovery, we need to combine the capabilities of the human and the machine as well as multiple methods and sources of evidence. Web Information Discovery Tool (WIDIT) Laboratory at the Indiana University School of Library and Information Science houses several projects that aim to apply this idea of multi-level fusion in the areas of information retrieval and knowledge organization. The TREC research group of WIDIT, who engages in examination of information retrieval strategies that can accommodate a variety of data environments and search tasks, participated in the Genomics, HARD, Robust, and Web tracks in TREC-2004. The basic approach of WIDIT was to leverage multiple sources of evidence, combine multiple methods, and integrate the strengths of man and the machine. Our main strategies for the tracks were: the use of gene name thesaurus in the Genomics track; query expansion and relevance feedback in the HARD track; query expansion with keywords from Web search in the Robust track, and the interactive system tuning process called "Dynamic Tuning" in the Web track.

2. Web track

In the Web track, we participated in the mixed query task as well as the query classification subtask. Our main strategies were fusion retrieval, where we combined different sources of evidence (e.g. body text, anchor text, header text), and post-retrieval reranking, where query type-specific methods were applied to adjust the document scores. The key component of WIDIT for the Web track was an interactive system tuning process called "Dynamic Tuning", which optimizes the fusion formula that combines the contributions of multiple sources of evidence (e.g. hyperlinks, URL, document structure). Dynamic tuning is a novel approach to system tuning that harnesses both the human intelligence and the computational power of the machine.

In addition to the dynamic tuning for fusion, we explored a query classification strategy that combines statistical and linguistic classification methods to identify the query type so that the system can adapt its retrieval methods according to the query type.

2.1 Query Classification

The goal of query classification task was to identify the categories of 225 mixed queries that consisted of 75 topic distillation (TD), 75 homepage finding (HP), and 75 named page finding (NP) queries. The main challenge of the query classification task stemmed from the short length of the queries, which contained only three words on the average (Table 1). We suspected that machine learning approaches may not be very effective in classifying texts with only a few words. Furthermore, the quality and the quantity of the training data available from previous years also seemed suboptimal for machine learning. There were 100 TD training queries compared to 300 HP and 295 NP queries, which were also short in length (Table 2) and appeared to be often ambiguous upon manual examination.

Table 1. 2004 Web track queries

Query Length (# of words)	1	2	3	4	5	6	7	Avg.
TD queries	12	41	16	6	0	0	0	2.2
NP queries	2	11	28	20	9	4	1	3.5
HP queries	3	9	38	15	8	2	0	3.3
All queries	17	61	82	41	17	6	1	3.0

Table 2. Training queries from 2001-2003 Web tracks

	Query Counts			Avg. Length (# of words)
	2001	2002	2003	
TD queries		50	50	2.8
NP queries		150	150	4.0
HP queries	145		150	3.6
All queries	145	200	350	3.7

Consequently, we decided to combine the statistical approach (i.e. automatic classifier) of machine learning with a linguistic classifier based on word cues. To supplement the training data for automatic classifiers, which had three times as many HP and NP than TD queries, we created a lexicon of US government topics by manually selecting keywords from the crawl of the Yahoo!'s U.S. Government category. The linguistic classifier used a set of heuristics based on the linguistic patterns specific to each query type identified from the analysis of the training data. For example, we noticed that queries that end in all uppercase letters tended to be HP, queries containing 4-digit year were more likely to be NP, and TD queries were shorter in general than HP or NP queries. We also identified some word cues for NP (e.g. about, annual, report, etc.) and HP (e.g. home, welcome, office, bureau, etc.) query types. After constructing the linguistic classifier, we combined the automatic classifier and the heuristic classifier with a simple ad-hoc heuristic that arrived at the query classification in the following manner:

if single word, assign TD.
else if strong word cue, assign linguistic classification.
else assign statistical classification.

We tested Naïve Bayes and SVM classifiers with the Yahoo-enriched training data, which showed little difference in performance. The classifier comparisons (i.e. statistical vs. linguistic vs. combination) showed the best performance by the combination classifier, which was the classifier used in our official run. Only three TREC groups who participated in the query classification task, and there was little difference in performance across systems.

2.2 Mixed Query Task

Our main task in the mixed query task was to optimize the system for mixed topic. Having engaged in the query classification, our approach to the mixed query task was based on optimizing retrieval strategy for each of the query types. To leverage the multiple sources of evidence, we created separate document indexes for body text, anchor text of incoming links, and header text that consists of meta field text and emphasized portion of body text. The retrieval

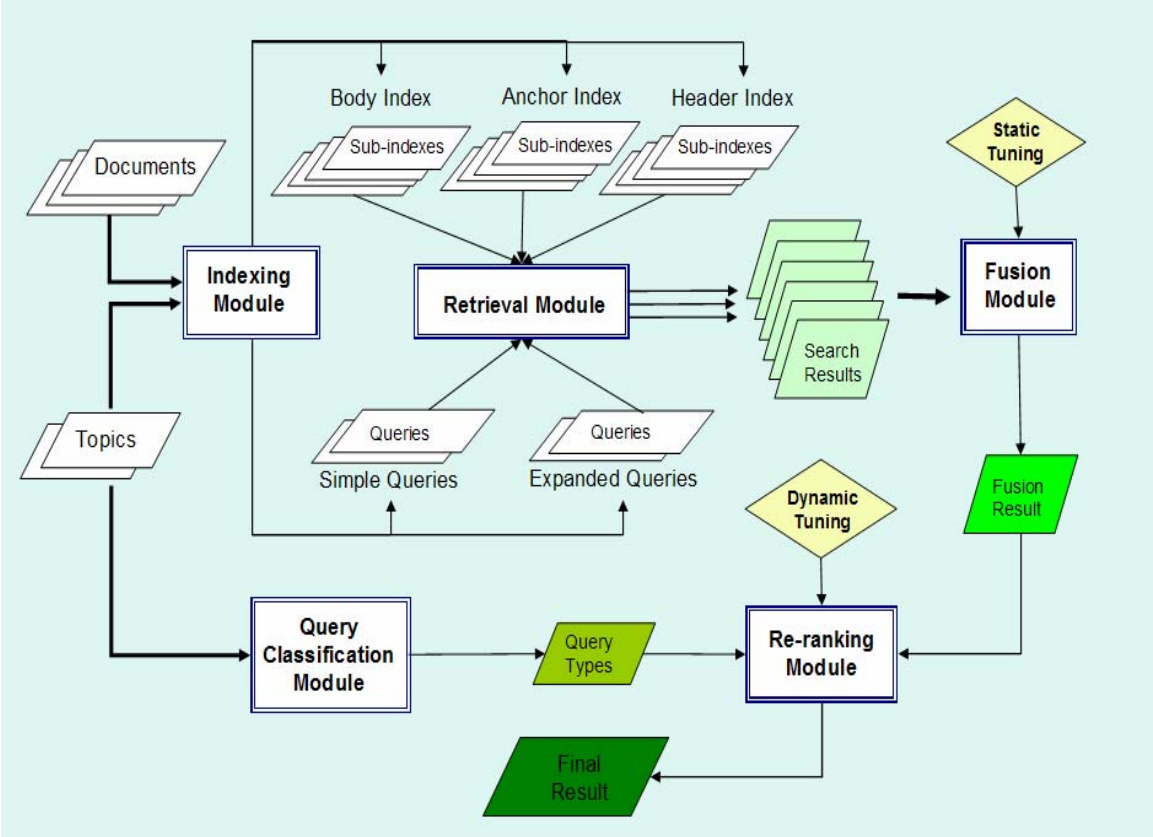
results using each index were combined using weighted sum with various weights to determine the optimum fusion formula for the baseline run without regards to query types.

In addition to fusion by result merging, we employed a post-retrieval rank-boosting strategy to rerank the merged results for each query type. Our general approach to query type-specific reranking was as follows: boost the rank of potential homepages if the query is topic-distillation or homepage finding type; boost the rank of pages with keyword matches if the query is hompage or named page finding type. More specifically, our rank boosting heuristic kept top 5 ranks static, while boosting the ranks of potential homepages (identified by URL type determination) as well as pages with keyword matches in document titles and URLs.

2.3 WIDIT Web IR System

WIDIT Web IR system consists of five main modules: indexing, retrieval, fusion (i.e. result merging), reranking, and query classification modules. The indexing module processes various sources of evidence to generate multiple indexes. The retrieval module produces multiple result sets from using different query formulations against multiple indexes. The fusion module, which is optimized via the static tuning process, combines result sets using weighted sum formula. The reranking module uses query-specific reranking formulas optimized via dynamic tuning process to rerank the merged results, and the query classification module uses a combination of statistical and linguistic classification methods to determine query types. The overview of WIDIT Web IR system architecture is displayed in Figure 1.

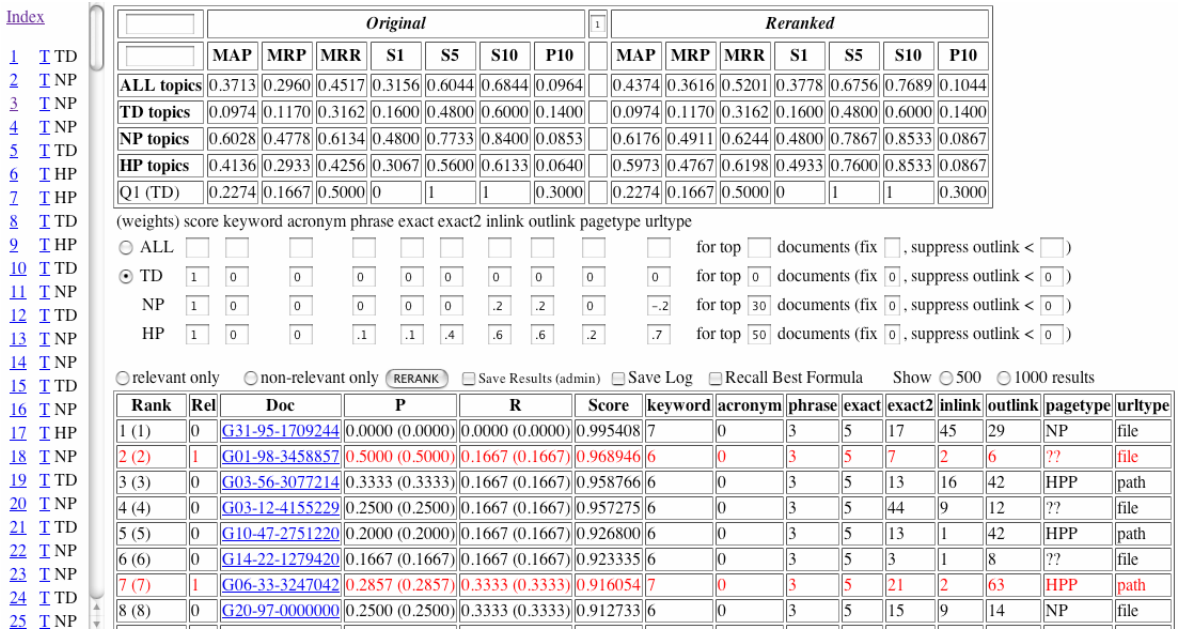
Figure 1. WIDIT Web IR System Architecture



2.3.1 Reranking Factors

Previous TREC participants found various sources of evidence such as anchor text (Craswell, Hawking & Robertson, 2001; Hawking & Craswell, 2002; Craswell & Hawking, 2003) and URL characteristics (Kraajj et al., 2002; Tomlinson, 2003, Zhang et al., 2003) to be useful in the Web track tasks. Based on those findings as well as the analysis of our Web track results in 2003, we decided to focus on four categories of the reranking factors. The first category was the field-specific match, where we scored each document by counting the occurrences of query words (keyword, acronym, phrase) in URL, title, header, and anchor texts. The second category of reranking factors we used was the exact match, where we looked for exact match of query text in title, header, and anchor texts (exact), or in the body text (exact2) of documents. The third category was link-based, where we counted documents' inlinks (indegree) and outlinks (outdegree). The last category was the document type, which was derived based on its URL (Tomlinson, 2003; Kraajj et al., 2002), or derived using a linguistic heuristic similar to the one used in query classification.

Figure 2. Dynamic Tuning Interface



2.3.2 Dynamic Tuning

Our findings from TREC-2003 (Yang & Albertson, 2003) indicated that fusion by result merging could be supplemented with post-retrieval reranking based on metadata (e.g. link count, URL characteristics) to enhance retrieval performance in the topic distillation task. In 2003, however, we were not successful in devising effective reranking strategies for the homepage and named page finding tasks, nor were we able to adequately address the question of how to deal with mixed query searches.

Thus, the focus of our TREC-2004 Web track efforts was to extend the fusion approach by introducing the "dynamic tuning" process with which to optimize the fusion formula that combines the contributions of multiple sources of evidence (e.g. hyperlinks, URL, document structure). The dynamic tuning process is implemented as a Web application (Figure 2); where interactive system parameter tuning by the user produces in real time the display of system

performance changes as well as the new search results annotated with metadata of fusion parameter values (e.g. link counts, URL type, etc.). The key idea of dynamic tuning, which is to combine the human intelligence, especially pattern recognition ability, with the computational power of the machine, is implemented in this Web application that allows human to examine not only the immediate effect of his/her system tuning but also the possible explanation of the tuning effect in the form of data patterns. By engaging in iterative dynamic tuning process, where we successively fine-tuned the fusion parameters based on the cognitive analysis of immediate system feedback, we were able to significantly increase our system performance.

The effective reranking factors observed from the iterations of dynamic reranking were: indegree, outdegree, exact match, and URL/Pagetype with the minimum number of outdegree of 1 for HP queries; indegree, outdegree, and URLtype for NP queries (1/3 impact of HP factors); acronym, outdegree, and URLtype with the minimum number of outdegree of 10 for TD queries. In addition to harnessing both the human intelligence and machine processing power to facilitate the process of system tuning with many parameters, dynamic tuning turned out to be a good tool for failure analysis. We examined severe search failure instances by WIDIT using the dynamic tuning interface and observed the following:

- Acronym Effect
  - WIDIT expanded acronyms and ranked documents about the acronym higher than the specific topic.
  - e.g. CDC documents ranked higher than Rabies documents for topic 89 ("CDC Rabies homepage")
- Duplicate Documents
  - WIDIT eliminated documents with the same URLs and ranked mirrored documents higher.
  - e.g. Relevant documents with the same URL (G00-74-1477693 and G00-05-3317821 for topic 215) were not indexed by WIDIT.
  - e.g. G32-10-1245341 is a mirror document of G00-48-1227124 (relevant for topic 188) but not counted as relevant by TREC official judgments.
- Link Noise Effect
  - Non-relevant documents with irrelevant links are ranked high by WIDIT
  - e.g. The relevant document for topic 197 ("Vietnam War") is Johnson Administration's "Foreign Relations" volumes with 4 links to Vietnam volumes, but WIDIT retrieved pages about Vietnam with many irrelevant (e.g. navigational) links at top ranks.
- Topic Drift
  - Topically related documents with high frequency of query terms were ranked high by WIDIT.
  - e.g. Documents about drunk driving victims, MADD, etc. were ranked higher than the impaired driving program of NHTSA page for topic 192 ("Drunk driving").

2.4 Web Track Results

Table 3 shows the results of our mixed query task runs. Since we were not able to fully implement the dynamic tuning module in time, our official submission consisted of fusion runs and reranking runs using a static reranking formula based on past findings regarding specific query types. The post-submission runs, which employed dynamic tuning, achieved better performance in general, especially when using the official query types. The best fusion run

combined the best baseline result, which used anchor text index, and the top two fusion runs, which merged the results of body, anchor, and header index results.

Table 3. Mixed Query Task Results (MAP = mean average precision, MRR = mean reciprocal rank)

	MAP (TD)	MRR (NP)	MRR (HP)
F3	0.0974	0.6134	0.4256
SR_g	0.0949	0.6018	0.4487
DR_g	0.1274	0.5418	0.6371
SR_o	0.0986	0.6258	0.4341
DR_o	0.1349	0.6545	0.6265
TREC Median	0.1010	0.5888	0.5838

F3: Best fusion run
SR_g: Static reranking run using the guessed query type
DR_g: Dynamic reranking run using the guessed query type
SR_o: Static reranking run using the official query type
DR_o: Dynamic reranking run using the official query type

In order to assess the effect of query classification error, we generated random assignment of query types (DR_r) and worst possible assignment of query types (DR_b). Table 4.1 compares the classification error of WIDIT query classification algorithm with random and worst classification. Because TD task is biased towards homepages, HP-TD error is the least severe type of error. Since HP and NP tasks are both known-item search task, HP-NP error is less severe than NP-TD, which is the least similar. In table 4.2, which shows the results of dynamic reranking using each query classification, we can see that random or poor query classification will adversely affect the retrieval performance. Table 4.2 also shows the random query type results to be comparable with TREC median performance for TD and HP queries.

Table 4.1 Query classification error by error type

	Error Type		
	HP-TD	HP-NP	NP-TD
DR_g	26	49	17
DR_r	54	48	44
DR_b		75	150

Table 4.2 Query classification error by error type

	MAP (TD)	MRR (NP)	MRR (HP)
DR_o	0.1349	0.6545	0.6265
DR_g	0.1274	0.5418	0.6371
DR_r	0.1235	0.4450	0.5285
DR_b	0.0922	0.2995	0.3105
TREC Median	0.1010	0.5888	0.5838

DR_o: Dynamic reranking run using the official query type
DR_g: Dynamic reranking run using the guessed query type
DR_r: Dynamic reranking run using the random query type
DR_b: Dynamic reranking run using the bad query type

3. HARD track

Conventional retrieval systems, which ignore the fact that users are different, often fail to satisfy the various aspects of user's information need beyond the topical relevance. The HARD (High Accuracy Retrieval from Documents) track, introduced in 2003, investigates approaches that can enhance the retrieval performance by tailoring the search to the user.

The HARD track has three phases. First, the HARD participant produces baseline retrieval results using the initial topic descriptions without any user-specific metadata. After the baseline run, the participant creates the Clarification Forms (CF), which is given to the user to collect relevance data for each query. In the third and final phase, the participant can use the metadata about queries (e.g. familiarity, genre, subjects, and geography) that are provided by the user, the relevance data collected from the clarification forms, or a combination of both to generate the final submission runs.

We participated in all three phrases (baseline run, clarification form, final run). The focal points of our strategy were query expansion and relevance feedback by clarification form. For the baseline retrieval, we examined the effectiveness of query formulation strategies with emphasis on automatic query expansion. Initial query formulations involved combinations of topic fields (title, description, narrative) and stemming (simple plural stemmer, combination stemmer), to which were added combinations of expansion components such as synonyms from the WordNet, noun phrases identified by Brill Tagger, expanded acronyms and word definitions from Web search. The analysis of the results based on the training data suggested that automatic query expansion with synonyms and word definition terms can introduce noise that hurt retrieval performance, whereas acronyms, nouns and noun phrases found in the topic titles tend to be terms with more discriminating power. We also investigated query expansion via pseudo-relevance feedback, but it showed adverse effect on retrieval performance.

In an attempt to decrease the noise introduced in automatic query expansion, we involved the user to filter the expanded query via clarification forms (Figures 3.1 and 3.2), where the user selected relevant query expansion terms and best sentences from top 25 baseline results for each query. The best sentence of a document was extracted using strategies from the past Q&A track. In the final run, the baseline query was to be modified with information from CF feedback as well as metadata information provided in the metadata version of HARD topics. Post-retrieval re-ranking and metadata labeling were the main tasks in this stage. In order to "label" (or extract) metadata on documents, different lexicon bases were generated for each metadata fields (e.g. location, subject) and documents were scored for each metadata using a combination of statistical and linguistic classification methods. Unfortunately, we were not able to implement the re-ranking module in time, which was to be based on explicit relevant judgment we got from CF as well as implicit clues, such as emphasis on specific fields, noun phrases, domain-specific lexicon use, and linguistic clues. Although our official submission included only the baseline runs and CF-enhance runs, we include the description of our original metadata strategy below.

3.1 Metadata Strategy

The four metadata types associated with HARD topics were geography, genre, familiarity, and subject. For geography, we created US and non-US location lexicon from mining the Web resources (e.g. Yahoo!) and counting the occurrences of the lexicon terms in the first line of news or keywords field. For genre, we considered the document with high proportion of quoted string as opinion/editorial. As for familiarity, we created a rare word lexicon from an online dictionary and scored documents by the proportion of rare words. We also created subject lexicon for each subject value by querying Yahoo! category and WordNet Hyponyms (… is a kind of subject) and counting the lexicon term occurrences in the keyword fields of the documents. The metadata

scores thus computed can then be used in the post-retrieval reranking process such as dynamic reranking.

Figure 3.1 Clarification Form I

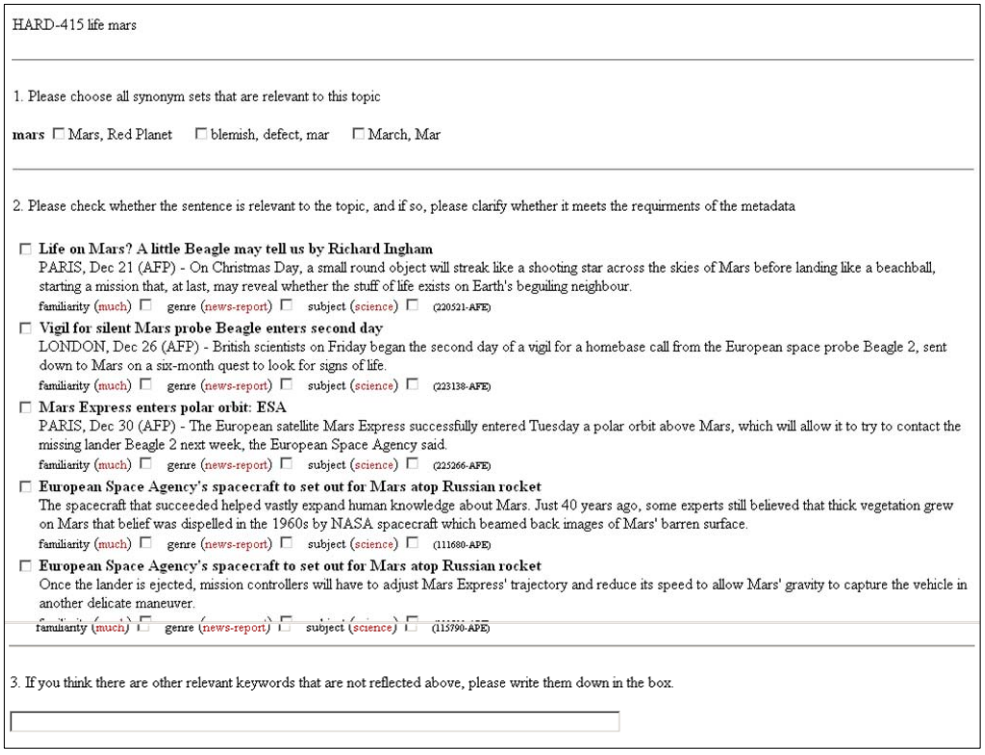
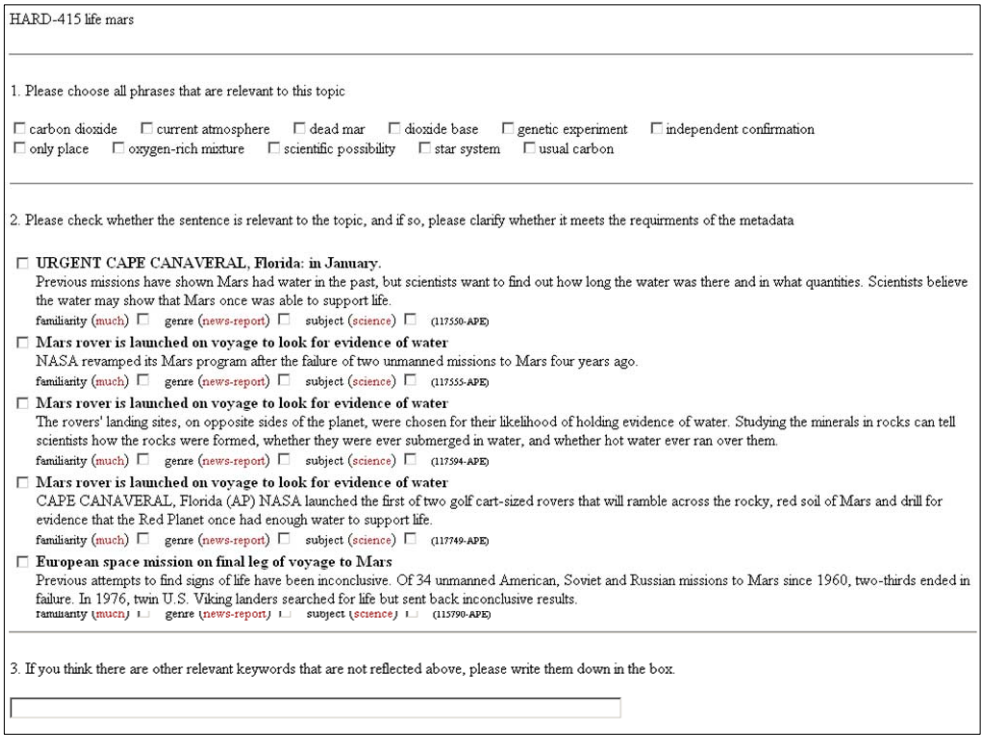


Figure 3.2 Clarification Form II



3.2 HARD Track Results

Table 5 shows our best baseline and CF result. The best baseline run, which used Okapi term weight, query expansion with acronyms and nouns, and the combo stemmer (Yang, Maglaughlin & Newby, 2001) that combines simple plural removal and inflectional stemming, performed well above the median level. The CF run, using the relevance feedback from the clarification form, improved the retrieval performance of the baseline only slightly, which suggests the effectiveness of the automatic query expansion in the baseline run.

Table 5. HARD results (MAP = mean average precision, MRP = mean R-precision)

	MAP	MRP
TREC Best	0.3554	0.3717
WIDIT CF run	0.3287	0.3454
WIDIT best baseline run	0.3128	0.3366
TREC Median	0.2634	0.2906
TREC Worst	0.0288	0.0673

4. Robust track

The Robust track explored methods for improving the consistency of retrieval technology by focusing on poorly performing topics. The 2004 Robust track was a classic ad-hoc retrieval task using 250 topics. Query expansion with keywords from Web search was the main WIDIT approach to the Robust track, which extended the methodology of the best Robust track system in previous TREC (Grunfeld et al., 2003).

We submitted four runs for the Robust track using combinations of topic field text, different term weighting formula, and query expansion methods. Title, description and narrative text were combined to create *wdoqla1* and *wdo25qla1*, but *wdoqla1* used the original Okapi term weight formula, whereas *wdo25qla1* used the modified Okapi BM25 formula. *wdoqdn1*, based on description field, and *wdoqsn1*, based on title field, expanded the query with nouns extracted by the Brill tagger. Both runs were weighted with the original Okapi term weights.

All runs used a simple affix removal stemming algorithm that included various topic-specific exception word lists. Stemmed words were then compared against a dictionary for accuracy. Other retrieval runs were attempted using query expansion with web search engines such as Yahoo, Google, Altavista and Search, as well as lexically-based query expansion methods with WordNet.com; however, these methods introduced a high level of noise and did not deliver good retrieval results and thus excluded from the official submission. Table 6, which shows the robust track results by topic type, indicates above median level of performance by WIDIT.

Table 6. Mean Average Precision of Robust runs by topic type

	Old Topics	New Topics	Difficult Topics	All Topics
TREC Best	0.3429	0.4227	0.1949	0.3586
WIDIT Best (*wdoqla1*)	0.2819	0.3300	0.1363	0.2914
TREC Median	0.2667	0.2979	0.1260	0.2755
TREC Worst	0.0692	0.0529	0.0207	0.0756

5. Genomics track

The Genomics track investigated how exploiting domain-specific information improves retrieval effectiveness. The 2004 genomics track contained an ad-hoc retrieval task and three variants of a categorization task using a 10-year subset (1994–2003) of MEDLINE data (4.5 million MEDLINE records, 9 GB) and 50 topics derived from information needs of biomedical researchers. One of the main WIDIT approach to the Genomics track was to build a gene name thesaurus by a combination of statistical (e.g. Latent Semantic Indexing) and linguistic (e.g. Gene Ontology harvest) clustering methods. We could not scale up the LSI module in time to handle the Genomics data, so we only used the gene synonyms created from the Gene Ontology harvest and nouns and phrases identified by the NLP module to expand the queries. For the Categorization task, we only attempted the triage task using a Naïve Bayes classifier. The WIDIT results for both ad-hoc and triage tasks were below the median level of performance.

References

Craswell,N., & Hawking, D. (2003). Overview of the TREC-2002 Web track. *Proceedings of the 11th Text Retrieval Conference (TREC 2002)*, 86-95.

Craswell, N., Hawking, D., & Robertson, S. (2001). Effective site finding using link anchor information. *Proceedings of the 24th ACM SIGIR Conference on Research and Development in Information Retrieval*, 250-257.

Grunfeld, L., Kwok, K.L., Dinstl, N., & Deng, P. (2003). TREC 2003 Robust, HARD, and QA track experiments using PIRCS. *Proceedings of the 12th Text Retrieval Conference (TREC2003)*, 510-521.

Hawking, D., & Craswell, N. (2002). Overview of the TREC-2001 Web track. *Proceedings of the 10th Text Retrieval Conference (TREC 2001)*, 25-31

Kraaij, W., Westerveld, T., & Hiemstra, D. (2002). The importance of prior probabilities for entry page search. *Proceedings of the 25th ACM SIGIR Conference on Research and Development in Information Retrieval*, 27-34.

Tomlinson, S. (2003). Robust, Web and Genomic retrieval with Hummingbird SearchServer at TREC 2003. *Proceedings of the 12th Text Retrieval Conference (TREC2003)*, 254-267.

Zhang, M., Lin, C., Liu, Y., Zhao, L., Ma, L., & Ma, S. (2003a). THUIR at TREC 2003: Novelty, Robust, Web and HARD. *The 12th Text Retrieval Conference (TREC 2003) Notebook*, 137-148.

Yang, K. & Albertson, D (2003). WIDIT in TREC2003 Web track. *Proceedings of the 12th Text Retrieval Conference (TREC2003)*, 328-336.

Yang, K., Maglaughlin, K., & Newby, G. (2001). Passage feedback with IRIS. *Information Processing & Management*, 37, 521-541.