# Rutgers Information Interaction Lab at TREC 2005: Trying HARD

N.J. Belkin, M. Cole, J. Gwizdka, Y.-L. Li, J.-J. Liu, G. Muresan, D. Roussinov*, C.A. Smith, A. Taylor, X.-J. Yuan
School of Communication, Information and Library Studies, Rutgers University
New Brunswick, NJ 08901-1071
*W.P. Carey School of Business, Arizona State University, Tempe, AZ 85287-4606
[belkin, mcole, jgwizdka, lynnlee, jingjing, muresan, csmith, xjyuan] @ scils.rutgers.edu
dmitri.roussinov@asu.edu, artaylor@eden.rutgers.edu

## Abstract

Within the structure of the TREC 2005 HARD track guidelines, we investigated the following hypotheses: **H1**: Query expansion using a "clarity"-based approach will increase effectiveness over baseline queries and baseline queries plus pseudo-relevance feedback; **H2**: Query expansion based on the Web will increase effectiveness over baseline queries and baseline queries plus pseudo-relevance feedback; **H3**: Query expansion using terms selected by the searcher from those suggested by clarity modeling and/or the web will increase effectiveness over baseline queries, baseline queries plus pseudo-relevance feedback, and queries expanded by all suggested terms; **H4**: Query expansion using "problem statements" elicited from the searcher will increase effectiveness over baseline queries and baseline queries plus pseudo-relevance feedback; **H5**: The effectiveness of query expansion using problem statements will be negatively correlated with query "clarity". H1 and H2 were tested without user intervention; H3 and H4 were tested using two different "clarification forms"; H5 was tested using the results of the H4 clarification form. Baseline queries were generated from the topic titles and descriptions; query expansion was accomplished by adding terms to the baseline queries, with a variety of weights given to the expansion terms, relative to the baseline terms. Preliminary results indicate that H1, H2, H3 and H4 are in part weakly supported, in that performance is increased over baseline, but it is not increased over pseudo-relevance feedback. H5 was not supported. Combining some degree of user interaction (H3) with pseudo-relevance feedback appears to lead to increased performance.

## 1 Introduction

The 2005 TREC HARD track differed from those in the previous two years in that there was no user "metadata", other than the assessor's degree of familiarity with the topic, all other information about the assessor's context having to be derived from interactive clarification forms (CFs). Since the CFs are understood to be simulations of interactions with a searcher, subsequent to an initial query, we were primarily concerned in this year's investigations with whether, and under what circumstances, such interaction is worthwhile. As has been demonstrated in numerous studies, searchers typically are willing to engage in such interaction only when the payoff, in terms of control and effectiveness, is perceived to be high, and the interaction is clearly relevant to the searcher's goal (e.g. Belkin, et al., 2000).

The primary purposes of the type of interaction represented by CFs have typically been understood to be either more detailed specification of the query/information problem, in particular for disambiguation; or indication of the searcher's context, used to tailor search results in some way or another. The former is typically implemented as some form of query expansion; the latter as re-ranking of the original search results. In both of these cases, the obvious

comparison to make with respect to the usefulness of the interaction is with the effectiveness of the original query; and with the effectiveness of the original query, modified by automatic methods which do not require interaction. Query expansion using pseudo-relevance feedback is an obvious candidate for a non-interactive method; so also are query expansion using an external resource, such as the Web, and using language-modeling/clarity rather than traditional relevance feedback methods for expansion term selection. In this year's research, we investigated the use of query expansion based on CFs, and compared performance under those queries with implementations of these two general automatic methods for query expansion, as well as with a baseline query derived only from the title and description of the topic. We did not consider the use of searcher's level of familiarity with the topic for search result modification.

We organized our investigations in HARD 2005 around several hypotheses, which, in standard scientific manner, we hoped to either disprove, or support. Our initial idea was to test whether searcher selection of terms thought by the system to be useful in query expansion was effective, as some studies have shown (e.g. Koenemann.and Belkin, 1996). We also wanted to investigate the effectiveness of query expansion terms under different selection models. We therefore implemented expansion term identification in three different ways: (i) standard pseudo-relevance feedback, as implemented in the Lemur toolkit[1]; (2) what we call "clarity scoring" of terms based on the database being searched, using Lemur utilities; and (3) use of the Web as an external resource, to identify terms and phrases associated with the query which might otherwise not be identified through database-specific methods. This led us to the first three hypotheses which we tested.

> **H1**: Query expansion using a clarity approach will increase effectiveness over baseline queries and baseline queries plus standard pseudo-relevance feedback
>
> **H2**: Query expansion based on the Web will increase effectiveness over baseline queries and baseline queries plus pseudo-relevance feedback.
>
> **H3**: Query expansion using terms selected by the searcher from those suggested by clarity modeling and/or the web will increase effectiveness over baseline queries, baseline queries plus pseudo-relevance feedback, and queries expanded by all suggested terms.

Our second area of concern was with elicitation of longer and more complex descriptions of the information problem than a typical keyword query such as that simulated by the content words of the title and description of a TREC topic. This approach is based on the general idea of the ASK hypothesis, that searchers should not be asked to specify that which they don't know (Belkin, 1980); on research results that indicate that longer queries in a best-match search system result in improved performance (e.g. Belkin, et al., 2002); and on results from the 2004 HARD track which showed performance improvement over baseline when such descriptions were used for query expansion (Kelly, Dollo & Fu, 2005). We speculated that elicitation of this type of "problem statement" might be useful for query expansion when an initial query was likely to be ineffective; following Cronen-Townsend, Zhou and Croft (2002), we operationalized ineffectiveness as low query clarity. This line of investigation led us to our other two hypotheses.

> **H4**: Query expansion using "problem statements" elicited from the searcher will increase effectiveness over baseline queries and baseline queries plus pseudo-relevance feedback.

[1] http://www.lemurproject.org/

**H5**: The effectiveness of query expansion using problem statements will be negatively correlated with query "clarity".

In the next section, we describe in detail how we implemented and used the clarification forms, and other data, for testing each of these hypotheses.

# 2 Clarification Forms

## *2.1 CF1: Internet Mining Terms*

Past experience with TREC topics indicates that, while the query expansion based on pseudo-relevance feedback (adding terms from the top returned documents to the query) is the most effective way of improving performance, it is not effective on the worst topics due to a phenomenon commonly known as "query drift:" when the top documents are irrelevant, so are the added terms.

Since this year's HARD track topics have been selected from the worst prior year topics, we were deliberately looking into different and hopefully safer expansion strategies. Inspired by the success of our (Roussinov & Zhao, 2003) and other researchers' (Kwok et al., 2004) work on Internet mining and user-controlled query expansion (Koenemann & Belkin, 1996), we developed a Navigation By Expansion (NBE) paradigm and tested it with our submitted runs. The NBE paradigm follows the same intuitive principles of navigation that people employ while, for example, driving or walking in new surroundings. First, the topic *surroundings* are identified (e.g. as a set of possibly related words or phrases) through the Internet mining (or possibly other resources, such as WordNet). Second, the set of possible *moves* within those surroundings is identified (e.g. by preserving only those terms that are present in the HARD collection).

Although the specific methods on which the paradigm builds (expansion, refining) have been studied in the past, we believe that *combining them into a single higher level framework is beneficial.* This formulation also influenced the particular choice of techniques and models to use in the implementation described below. We also believe that the approach has not been methodologically studied yet in spite of its promise. For example, it is still not known what specific techniques work best for representing the topic surroundings, what models work best for selecting candidate terms for the query expansion, and what are the most effective ways of ranking the documents while using the expanded query.

In this year's TREC, in order to determine the surrounding concepts, we submitted our query to Google and built what we call an *Internet language model* for it. Only the concepts with the frequencies of occurrence among the top 200 returned pages (full-text) larger than their background frequencies of occurrence on the Web were considered. We designed and trained a special formula for the probability of being a surrounding concept using logistic regression and different topics from the previous years. Some additional details on the mining approach and its applications beyond ad hoc retrieval can be found in Roussinov et al. (2005). Our set of CF1 forms implemented the proposed NBE paradigm.

The Internet language model (probabilities of occurrences of possibly related terms) for each topic was built and analyzed through the following steps:

Step 1. The title and description were merged into a single query and sent to Google.

Step 2. The full text of the top 200 pages returned by Google was downloaded as the *mining corpus* (the "ore").

Step 3. For each term (a sequence of up to 3 consecutive words) in the mining corpus, the probability of being "related to the topic" was estimated by approximating the logistic regression on the deviation from randomness when the values of the probability were approaching 1, specifically as following:

$$Pr(t) = 1 - \exp(-(s - 1) / \alpha), \text{ where:}$$

s = signal to noise ratio of the term, estimated as:

$$s = (df_m / N_m) / (df_w / W), \text{ where}$$

$df_m$ was the number of occurrences of the term in the mining corpus,
$N_m$ was the number of pages in the mining corpus,
$df_w$ was the number pages on the Web in which the term occurs, obtained by querying Google,
W was the total number of pages covered by Google, set at 3,000,000,000 at the time.

$df_m / N_m$ represented "signal", while $df_w / W$ represented the "noise." For the non related term, we would expect the proportion of the pages within the mining corpus that have this term to be the same as the proportion of the pages having this term on the entire Web. The ratio of those two proportions represented the deviation from randomness within the mining corpus.

The adjustment parameter $\alpha$ defined how "steep" the probability curve was relatively to the signal to noise ratio. We set $\alpha$ to 0.5 by visually inspecting the related concepts for the different topics from the preceding years. With this value, a signal to noise ratio of 1.5 would give the probability of $1 - \exp(-1) = .63$. A signal to noise ratio of 2.5 would result in p = .86, etc. In our case, the results were not very sensitive to the value of parameter $\alpha$ since we only needed to select the top most deviant from the background terms, so we did not rely on the actual probability estimate. Although including the terms that are frequent in the mining corpus as a result of their being frequent in the entire Web would not generally hurt the retrieval results since they would typically have low *idf* value, it would still put higher cognitive load on the user and require more time to make selections. Since our NBE paradigm requires the expansion terms to represent concept surroundings, the value of Pr(t) served as a good guidance which terms to use in the clarification forms.

Step 4. Out of the related terms, we preserved only those that occurred at least once in the target collection (Aquaint) and sorted them according to their *impact estimate*:

$$i = Pr(t) * idf, \text{ where}$$

idf was the inverse document frequency computed based on the target collection statistic: $idf = \log(N/df) / \log(N)$, where N was the number of documents in the HARD corpus and df was the number of documents containing the term t. Since idf weight was used by our document ranking function (BM25), it provided a reasonable estimate of each term impact (if selected) relatively to the other selected terms. Using the terms with low idf values in the clarification would not hurt the performance, but again would not be an efficient "move" within our NBE paradigm since their effect on ranking would be negligible.

### 2.2 CF1: Clarity Terms

The second set of terms that we used was created based on the notion of query *clarity* (Cronen-Townsend, Zhou and Croft, 2002). The clarity of each query, and each of its individual terms was computed by the Lemur toolkit that we used for indexing and ranking (QueryClarity application). We used the following (default) parameters: feedbackDocuments obtained by the BM25 retrieval from the target collection (also with the default parameters), feedbackDocCount = 5, feedbackCoefficient = .5, feedbackTermCount = 100. We sorted the terms according to the clarity values reported by Lemur and selected the top 10. The Lemur QueryClarity application does not support phrases, thus all our clarity terms were single word terms. Since all the selected terms were from the target collection, they all represented valid "moves" within our NBE paradigm.

### 2.3 CF1: Merging Terms

To produce CF1 for presentation to the assessors, we merged the top ten terms from the two different sources, removed duplicates, and presented them to the assessors in alphabetical order, with no reference to the source of the terms. Assessors were asked to choose those terms which they thought would be useful in expanding the original query (see Figure 1). Each form listed the topic title, description and narrative, followed by up to 20 terms (words or phrases), each preceded by a check box. The instructions to select the term were the following: "The search system has identified several search terms which are thought to be possibly useful for modifying the search query for this topic. Please check all of the search terms which you think would be good to include in a query for this topic." When the user selected the box, the term was used in subsequent query expansion, otherwise it was ignored.

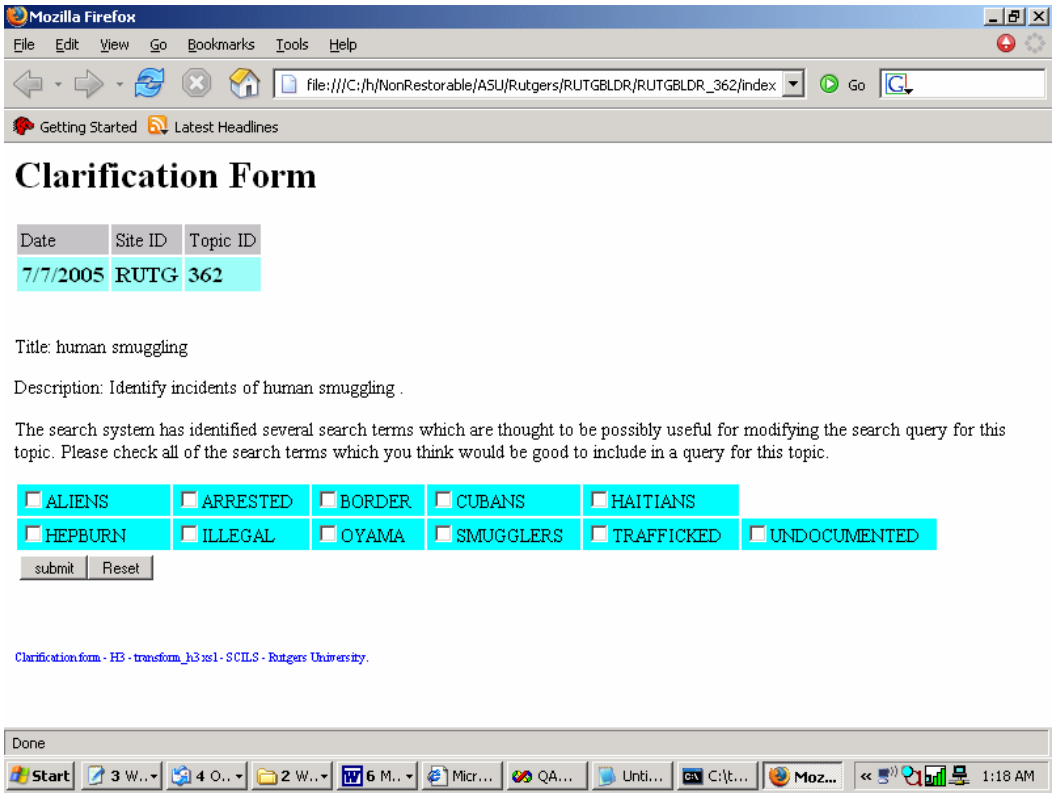


**Figure 1. The clarification form (CF1) for the topic "Human Smuggling."**

### *2.4 CF2: User Generated Terms*

Following our work in the Interactive Tracks of TREC 2002 (Belkin, et al., 2002) and 2003 (Belkin et al., 2003), our fourth hypothesis was that retrieval performance would be improved if additional terms generated by the user were added to the query. Following the experiment reported by UNC for TREC 2004 (Kelly, Dollu, & Fu, 2004; 2005) we used a second clarification form (CF2), which presented three specific requests for additional terms. In the UNC study, which also presented three open-ended elicitations, the order of the requests was not rotated among subjects. UNC found that the response to their first request, "Describe what you already know about this topic," produced more terms on average than the other two subsequent questions (30.98 vs. 23.11 and 2.47 terms respectively). In addition, when the performance of each of the three elicitations was compared separately to baseline, the first request had the largest positive effect on performance ($p < .05$) (Kelly, Dollu, & Fu, 2005). The UNC papers suggested that the relative superiority of their first open-ended request might be explained as an order effect. One of our objectives was to explore whether information about background knowledge, as requested in the first UNC prompt, was more valuable than other sources of additional contextual terms, as requested in the second and third UNC prompts.

Our CF2 clarification form was designed to replicate the UNC experiment with control for elicitation order. For this reason, we rotated the order of the three open-ended prompts in our 50 clarification forms. Two of the three UNC elicitations were used in our CF2 form; in addition to the above first elicitation we also used UNC's third elicitation, "Please input any additional keywords that describe your topic." The second UNC elicitation was a question asking the subject "Why do you want to know about this topic?". Because subjects for HARD 2005 were not working on topics they created, which was the case in 2004, this second question did not apply. We replaced this question with one taken from the interview question used in the original ASK study (Belkin, Oddy, & Brooks, 1982, p. 146), asking the subject, "What sort of information would you like to have as a result of this search?". In the original ASK study, as in the current experiment, the assessment task was undertaken by an agent for the principal information seeker. For this reason we considered our substitute question to be a reasonable replacement for UNC's question related to the expected value of the output of the search.

## 3 Official Runs Description

### *3.1 Basic Runs*

Our basic approach was to supplement a query based on the title and description of the topic with terms drawn from two types of sources. The first type depended upon identification of terms in text collections. The other type depends on the assessor's understanding of their task and knowledge of the topic. For both types of information the additional terms were used for query expansion and document retrieval.

We used the Lemur IR toolkit to build the queries and retrieval results. With our interest in the effect of combining additional terms from several different types of sources, including user interaction, the structured query evaluation module (StructQueryEval), based on the InQuery retrieval model was employed. The results were evaluated using the standard trec_eval program (http://trec.nist.gov/trec_eval/).

Our original baseline run, officially submitted to NIST, was deficient in some way that we have not yet identified (MAP of 0.06 seems very unlikely), and so we replaced it in our set of official runs with another baseline run (RUTGBL) which was constructed as originally intended, using

the title and description fields of the test topics. RUTGBL is used as the baseline for evaluation of our results. Another submitted run (RUTGBF3) was based on pseudo-relevance feedback. Another run combined all the sets of expansion terms (from the Web, from Lemur, based on clarity, and derived from the answers solicited from assessors in the clarification form) with the topic title and description (RUTGALL).

The other runs were based on structured queries as a weighted sum of the topic and title with combinations of the various term sets. Three of those runs concerned combinations of the external term sources without assessor interaction (RUTGWS1 – expanded with terms derived from the web, RUTGLS1 – expanded with "clarity" terms, RUTGAS1 – expanded with both term sources). Two runs concerned explicit assessor interactions with the clarification forms (RUTGUS1 – terms selected by the assessor from CF1, RUTGUG1 – terms provided by the assessor in CF2). The details of the query formulation are given in Table 1.

The RUTGRS1 run was intended to capture a random sample of terms from those presented to the assessor in CF1. With proper re-sampling the performance of user selection can be compared with automatic random selection of the same number of terms. No meaningful conclusions can be drawn, however, from the single run and so there is no further discussion of this run here.

**Table 1: Query run construction**

| ***run*** | ***title*** | ***description*** | ***web*** | ***clarity*** | ***Presented in CF1[2]*** | ***Selected from CF1*** | ***Q1[3]*** | ***Q2[4]*** | ***Q3[5]*** |
|---|---|---|---|---|---|---|---|---|---|
| *RUTGBL* | ***1*** | ***1*** | | | | | | | |
| RUTGBF3[1] | 1 | 1 | | | | | | | |
| RUTGALL | 1 | 1 | 1 | 1 | | | 1 | 1 | 1 |
| RUTGAS1 | 0.9 | 0.9 | | | 0.1 | | | | |
| RUTGWS1 | 0.9 | 0.9 | 0.1 | | | | | | |
| RUTGLS1 | 0.9 | 0.9 | | 0.1 | | | | | |
| RUTGUS1 | 0.9 | 0.9 | | | | 0.1 | | | |
| RUTGUG1 | 0.9 | 0.9 | | | | | 0.1 | 0.1 | 0.1 |

The numbers in each cell indicate the relative weight given to terms from each of the term sources.
[1] Pseudo-relevance feedback with 20 documents and 100 terms, feedback coefficient=0.3
[2] The terms in CF1 were the Web and Lemur suggested terms with duplicates removed
[3] "Please describe what you already know about this topic."
[4] "What sort of information would you like to have as a result of this search ?"
[5] "Please input any additional keywords that describe this topic."

There were several problems implementing these intentions in our official runs. In several cases, due to a script bug, the CF1 clarification forms did not provide a few of the unique terms in the web and clarity suggested term sets. In the case of one topic (689) there were no clarity terms available, and for five topics in the web-suggested terms (435, 439, 443, 448, and 622) the queries were incorrect. We report results here with the deficiencies addressed in the affected web-suggested runs (RUTGALL and RUTGWS1).

There was also a problem with the intended weighting scheme for our runs, due to our inexperience in using the structured query syntax employed by Lemur. The weights were applied to the term sets taken as a group, rather than to the individual terms. So the official runs were

based on the relative contribution of the terms in each group, irrespective of the numbers of terms in each. Our intention was to weight individual terms drawn from the appropriate terms sources in the scheme given in Table 1. In the corrected runs the second weighting scheme has been adopted.

### *3.2 Combination Runs*

Since we were also curious to see how our NBE approach can work in combination with the pseudo-relevance feedback, we created two more runs using the original (non-expanded) baseline, expanded with pseudo-relevance feedback. We used the default parameters in Lemur to create it: feedbackDocCount = 20, feedbackTermCount = 100, feedbackPosCoeff = .3. This created a *pseudo-relevance feedback score*. Then, we implemented our own expansion module using the available C++ source in Lemur package. The top 1000 documents from the baseline were re-ranked according to the following score:

score = pseudo-relevance feedback score + 0.3 * expansion score,

where the expansion score was obtained using BM25 ranking with the default parameters for the query consisting only of the user selected CF1 terms. We used an expansion factor of 0.3 instead of 0.1 used with the other runs deliberately to diversify our set of official runs. This run is designated RUTBE in our results. The second combined run, RUTIN, was obtained using structured query in Lemur (StructQueryEval) and all the Web suggested terms added with 0.3 factor. We also specified Lemur to use blind feedback while doing structured query retrieval, with the following parameters: feedbackDocCount = 5, feedbackTermCount = 5, feedbackPosCoeff = 0.3.

## 4 Results

### *4.1 Our runs versus overall median runs.*

We start by comparing our baseline run (RUTGBL) with the overall baseline median results. R-precision and mean average precision of our baseline run were better than the overall baseline median (Table 2). When compared on individual topics, R-precision of our baseline run was better than the overall baseline median results on 25, worse on 18, and tied on 7 topics. However, these differences were overall not statistically significant.

**Table 2. Our baseline run compared with the overall baseline median results.**

| | R-precision | | MAP | | p@10 | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Overall Baseline median | 0.252 | 0.149 | 0.190 | 0.147 | 0.408 | 0.28 |
| RUTGBL (baseline) | 0.270 | 0.167 | 0.206 | 0.163 | 0.408 | 0.30 |

Next we compare our best run (RUTGALL) with the overall final median results (Table 3). R-precision and mean average precision of our RUTGALL run were statistically significantly better than the overall final median (Wilcoxon tests were: Z=-2.54, p=.011 and Z=-2.29; p=.003 respectively). R-precision of our RUTGALL run was better than the overall final median on 29, worse on 19, and tied on 2 topics. RUTGALL p@10 was better, but the difference was not significant.

**Table 3. Our best run compared with the overall final median results.**

| | R-precision | | MAP | | p@10 | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Overall Final median | 0.264 | 0.152 | 0.207 | 0.161 | 0.45 | 0.30 |
| RUTGALL | **0.299*** | **0.182** | 0.253 | 0.188 | **0.49**** | **0.31** |

### *4.2 Comparison of our experimental runs.*

The results of our experimental runs are presented in Table 4 and in Figure 2.

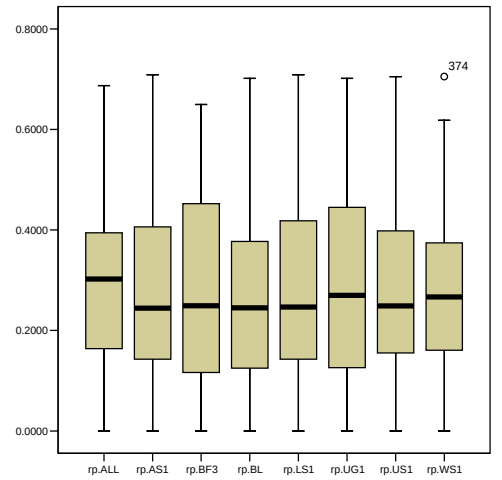


**Figure 2. R-precision for all**

R-precision and p@10 were not significantly different for all runs considered together (assessed by Friedman test for k-related samples: $\chi^2$=11.32, p=.125, $\chi^2$=9.60, p=.212, for R-precision and p@10 respectively). The mean average precision was significantly different ($\chi^2$=36.64, p=<.0001). Hence, we cannot establish an absolute ordering of all results in terms of their performance as measured by R-precision. However, based on pair-wise comparisons (Wilcoxon test) we can state that runs which incorporated all of the suggested terms (both clarity and web) (RUTGAS1), clarity suggested terms (RUTGLS1), web suggested terms (RUTGWS1), user selected terms (RUTGUS1), and user generated terms (RUTGUG1) were all individually significantly better than our baseline run, thus partially confirming hypotheses: H1,H2, H3, H4. Due to high variation of R-precision across topics, we cannot state whether blind feedback (RUTGBF3) and all suggested and user generated terms (RUTGALL) runs were significantly better than the baseline. Comparing RUTGALL to RUTGBL visually (Figure 3) we can see that on more than half of the topics (27/50) RUGTALL was better than or equal to (2/50) the baseline run.

**Table 4. Comparison of our runs against the baseline run (RUTGBL). The runs are sorted by ascending mean. Z-score and p-values were assessed using Wilcoxon signed ranks test.**

| Run name | R Precision | | | | Precision at 10 | | | | Mean Average Precision | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Z | p | Mean | SD | Z | p | Mean | SD | Z | p |
| **RUTGBL** | 0.270 | 0.167 | n/a | n/a | 0.408 | 0.30 | n/a | n/a | 0.206 | 0.16 | n/a | n/a |
| **RUTGAS1** | 0.278 | 0.168 | -3.01 | **0.003** | 0.430 | 0.31 | -2.21 | **0.027** | 0.216 | 0.17 | -4.25 | **0.000** |
| **RUTGLS1** | 0.279 | 0.169 | -1.99 | **0.046** | 0.424 | 0.32 | -1.29 | 0.196 | 0.218 | 0.17 | -3.36 | **0.001** |
| **RUTGWS1** | 0.281 | 0.166 | -2.96 | **0.003** | 0.426 | 0.31 | -1.88 | 0.060 | 0.219 | 0.17 | -4.04 | **0.000** |
| **RUTGUS1** | 0.282 | 0.166 | -2.49 | **0.013** | 0.440 | 0.31 | -2.43 | **0.015** | 0.222 | 0.17 | -4.35 | **0.000** |
| **RUTGUG1** | 0.286 | 0.173 | -2.93 | **0.003** | 0.458 | 0.31 | -2.78 | **0.005** | 0.228 | 0.17 | -4.96 | **0.000** |
| **RUTGBF3** | 0.287 | 0.206 | -0.90 | 0.368 | 0.480 | 0.36 | -1.76 | 0.078 | 0.248 | 0.22 | -2.23 | **0.026** |
| **RUTGALL** | 0.299 | 0.182 | -1.51 | 0.132 | 0.494 | 0.31 | -2.37 | **0.018** | 0.253 | 0.19 | -2.58 | **0.010** |

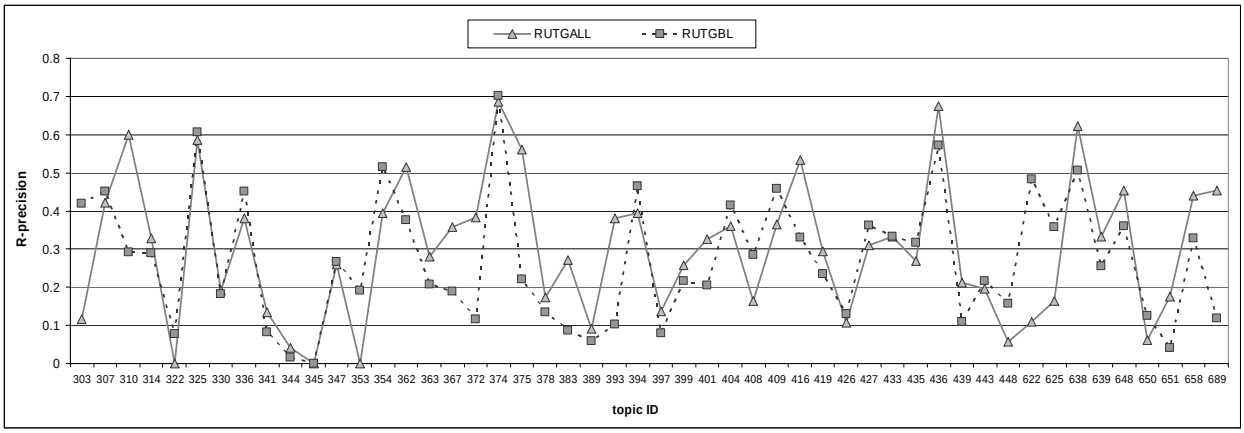


**Figure 3. Performance (R-precision) of RUTGALL versus RUTGBL on individual topics.**

We also performed pair-wise comparisons of our experimental runs with the blind feedback run (RUTGBF3). Blind feedback performance was not significantly different from other runs, with an exception of mean average precision, which was significantly better than for the baseline run (RUTGBL, Wilcoxon test: Z=-2.23; p=0.026). Thus our hypotheses H1,H2, H3, H4 cannot be confirmed with respect to performance of runs employing blind feedback in addition to baseline queries.

**Table 5. Summary of significant and non-significant differences in R-precision among our experimental runs.**

| | BL | AS1 | LS1 | WS1 | US1 | UG1 | BF3 | ALL |
|---|---|---|---|---|---|---|---|---|
| **BL** | ---- | | | | | | | |
| **AS1** | > | ---- | | | | | | |
| **LS1** | > | n/s | ---- | | | | | |
| **WS1** | > | n/s | n/s | ---- | | | | |
| **US1** | > | n/s | n/s | n/s | ---- | | | |
| **UG1** | > | n/s | n/s | n/s | n/s | ---- | | |
| **BF3** | n/s | n/s | n/s | n/s | n/s | n/s | ---- | |
| **ALL** | n/s | n/s | n/s | n/s | n/s | n/s | n/s | ---- |

> = row significantly better than column at $p \leq 0.05$, pairwise Wilcoxon signed ranks test

### *4.3 Analysis of CF2-based runs*

While we did not submit separate runs for each of the three elicitations in our second clarification form (CF2) we did produce unofficial runs for each (Q1, Q2, Q3), and for three combinations of the elicitations (Q1Q2, Q1Q3, Q2Q3, ), as well as the official run combining all three (RUTGUG1). We first compared the relative effects of the original query and CF2 terms where one component was included with varied weights, while the other with weight held constant at 1.0. We then focused on the effects of elicitation (Q1Q2, Q1Q3, Q2Q3, and Q1Q2Q3) and compare them with each other for runs with equal weighting of CF2 terms and the original query. We performed the following two sets of runs with varied weights:
1) original query (title + description) varied with weights ranging from 0.0 to 1.0, with the target model (i.e. the user-generated terms from the various questions) held constant with weight=1.0;

2) original query (title + description) held constant with weight=1.0, while the weight of the target model varied between 0.0 and 1.0. Figures 4 and 5 present the results of these runs.

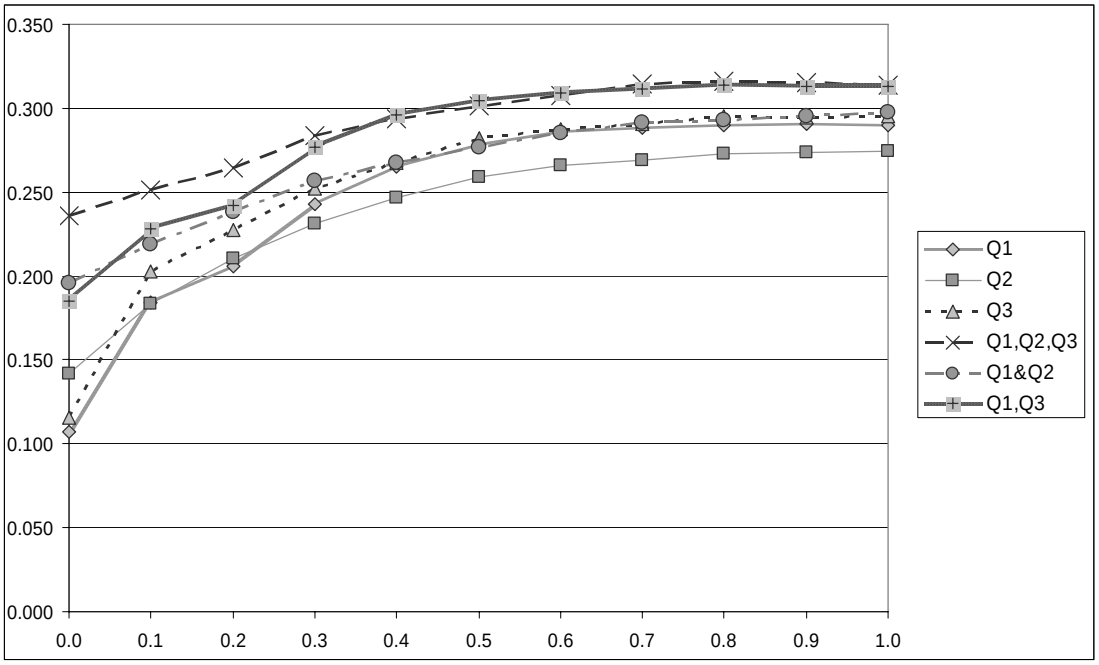


**Figure 4. R-precision (mean) for the original query (title + description) with variable weights 0.0-1.0, with user-generated terms at constant weight of 1.0**

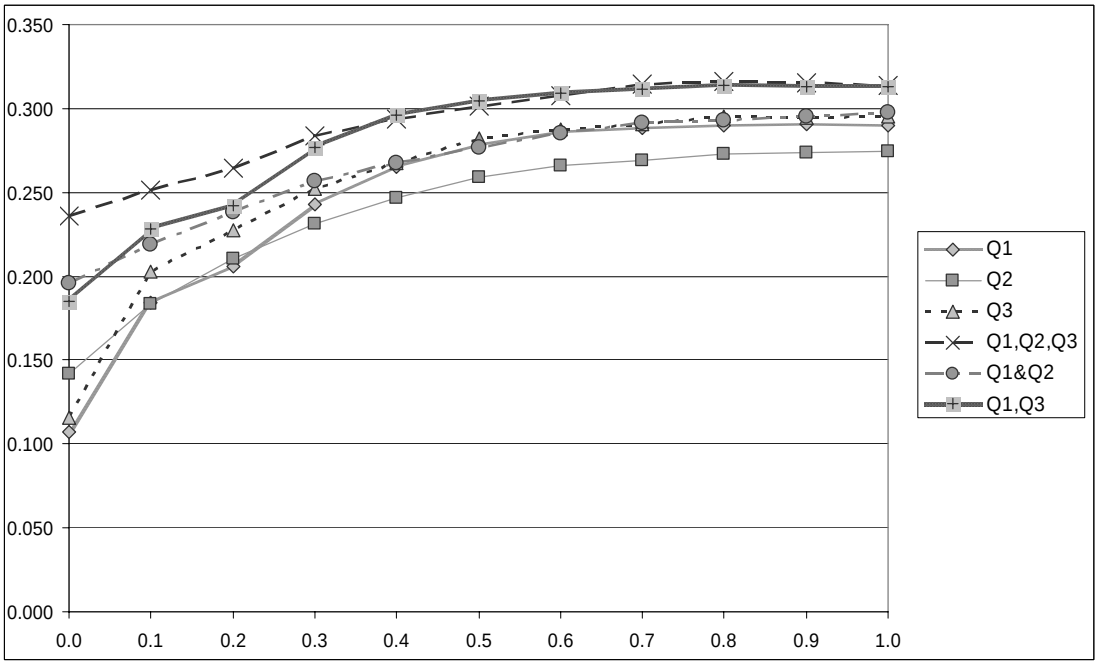


**Figure 5. R-precision (mean) for user-generated terms with variable weights: 0.0-1.0 and constant weight of 1.0 for the original query (title + description).**

Increasing the weight of the original query with respect to the user-generated terms significantly improved performance. R-precision was significantly better for runs with successively increased

weights (in all cases $\chi^2$>100, p<.001; Friedman test) (Figure 4). While R-precision also improved with increased weights of CF2 terms, the improvement over the original query was smaller, and only in three cases significant. The improvement was significant for: Q1, Q1&Q3 and Q1,Q2,Q3 (Friedman test: $\chi^2$=26.73, p=0.003, $\chi^2$=21.28, p=0.019, and $\chi^2$=30.833, p=0.001 respectively).

From figures 4 and 5, it is clear that giving equal weight to the original query terms and to the terms generated by the user in CF2 is an optimum strategy, and that our official run RUTGUG1, in which CF2 terms were combined with the original query in the ratio of 0.1/0.9 is not the best choice. This is confirmed by the data in table 6, in which equal weighting of all CF2 terms with the original query terms leads to performance substantially (although not statistically significantly) better than for RUTGUG1 (R-Precision 0.314 vs 0.286).

In order to understand more about factors related to these results, we had two objectives. The first was to determine the relative contribution of each of the three elicitations in the performance of our submitted RUTGUG1 run. We also wanted to compare our results, which were controlled for the order in which assessors answered the three elicitations, with UNC's results from TREC HARD 2004, where the elicitations were not controlled for order. These two objectives are discussed below.

**Table 6. CF2 Comparison with Revised Baseline and Runs with Terms from Three Elicitations and Three Combinations of Terms** (all runs used equal weighting of original query and user generated terms)

| Run name | R-Precision | | Precision at 10 | | Mean Average Precision | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| **RUTGBL** | 0.270 | 0.167 | 0.408 | 0.3 | 0.206 | 0.16 |
| **Q1** | 0.290 | 0.178 | 0.498* | 0.325 | 0.236 | 0.183 |
| **Q2** | 0.274 | 0.181 | **0.474*** | **0.321** | 0.223 | 0.181 |
| **Q3** | 0.295 | 0.164 | **0.498**** | **0.303** | **0.237**** | **0.175** |
| **Q1Q2** | **0.298*** | **0.182** | **0.514**** | **0.326** | **0.248**** | **0.190** |
| **Q1Q3** | **0.313*** | **0.176** | **0.538***** | **0.314** | **0.263**** | **0.186** |
| **Q1Q2Q3** | **0.314**** | **0.179** | **0.564***** | **0.304** | **0.268**** | **0.190** |

Q1 – "*Please describe what you already know about this topic."*
Q2 – "*What sort of information would you like to have as a result of this search?"*
Q3 – "*Please input any additional keywords that describe this topic."*

As can be seen in table 6, when used independently, all three elicitations produced terms that enhanced performance at the top of the list, as measured by p@10 (Q1: Z = -2.415, p < .05; Q2: Z = -2.117, p < .05; Q3: Z = -3.132, p < .01), while only Q3 produced significantly enhanced performance over baseline across the entire list, as measured by MAP (Z = -2.241, p < .05).

The best performing combination used terms from all three sources (Q1Q2Q3), however, this run was not significantly better than the run including only terms from Q1 and Q3 (see table 7), while it is significantly better than the Q1Q2 run. This finding raises the question of whether the Q2 terms are contributing to performance. The next logical question is the relative value of Q1 and Q3. A comparison of the combined Q1Q2 with the full Q1Q2Q3 run revealed that the addition of Q3 terms produced significantly enhanced performance. Clearly the Q3 terms are valuable, however, when the combined Q1Q2 run was compared with the Q1Q3 run, no significant performance difference was detected. As is revealed in the comparison of Q1 with

Q1Q3, the addition of Q3 terms did not produce significantly enhanced performance at the top of the list, but did so across the entire list (as measured by R-precision and MAP The comparison of Q3 and Q1Q3 runs revealed no significant difference in performance. Two conclusions are suggested by the above results. First, the value of the additional Q2 terms may not justify the costs (e.g. cognitive processing) of producing them. This point is related directly to discussion of UNC's results, below. Second, the Q3 elicitation may be producing the most valuable terms. This finding does not, however, indicate that Q3 would be the most productive elicitation in a more natural setting. Assessors completed our CF2 forms while completing forms from other participants. Q3 asked for additional keywords. It is possible, perhaps even likely, that assessors returned terms they remembered from prior forms. Further research requires more controlled experimentation.

**Table 7. CF2 Selected Comparison of Combinations of Terms from Three Elicitations (all runs used equal weighting of original query and user generated terms)**

| RUNS | r-precision | precision@10 | MAP |
|---|---|---|---|
| Q1 vs. Q1Q3 | **0.290 vs. 0.313 p < .05** | 0.498 vs. 0.538 n/s | **0.236 vs. 0.263 p < .01** |
| Q3 vs. Q1Q3 | 0.295 vs. 0.313 n/s | 0.498 vs. 0.538 n/s | 0.237 vs. 0.263 n/s |
| Q1Q2 vs. Q1Q3 | 0.298 vs. 0.313 n/s | 0.514 vs. 0.538 n/s | 0.248 vs. 0.263 n/s |
| Q1Q2 vs, Q1Q2Q3 | 0.298 vs. 0.314 n/s | **0.514 vs. 0.564 p < .01** | **0.248 vs. 0.268 P < .05** |
| Q1Q3 vs. Q1Q2Q3 | 0.313 vs. 0.314 n/s | 0.538 vs. 0.564 n/s | 0.263 vs. 0.268 n/s |

We have not yet completed the analysis of order effects, so it is not possible to draw conclusions from a comparison with UNC's prior results. One result is worth noting, however. Recall, Q2 was a reformulated question for our TREC HARD 2005; the question was designed to replicate UNC's question related to the expected value of the output of the search (Kelly, Dollu, & Fu, 2004; 2005). In both the results reported above, and in UNC's prior results, when used independently of the other two elicitations, this question produced the least valuable terms among the three elicitations. Without the order effect analysis, it is not possible to draw conclusions about this; however, it does suggest that the form of an elicitation may affect the retrieval value of the terms produced. This question is deserving of further research.

### *4.4 Testing H5*

Using the standard Lemur QueryClarity application, we tested to see whether there was any relationship between the effectiveness of query expansion using CF2 and initial query clarity of the topics. Tests were carried out using the following measures of improvement: percentage improvement; absolute magnitude of improvement; and, direction of change. We looked both for correlation of improvement with clarity score, and for any clarity score cut-off value which would identify topics for which CF2 intervention would be valuable. On all measures, and for both cases, we could not identify any significant relationships.

### *4.5 Combining CF1 results with pseudo-relevance feedback*

We investigated one other aspect of the use of our CF1 terms. Given that modifying the original query (RUTGBL) with pseudo-relevance feedback (RUTGBF3) led to somewhat improved

performance, we were interested in whether the pseudo-relevance feedback performance could be further improved by additional sources of evidence. We tested this issue with the two "combination" runs described in section 3.2. RUTBE, a run which augmented the original query plus pseudo-relevance feedback with the user-selected terms from CF1 with a relative weight for the last of 0.3, performed significantly better than the baseline, RUTGBL, and also better than RUTGBF3, on all measures. However, the RUTBE used BM25 weighting, whereas RUTGBL and RUTGBF3 used InQuery weighting, so these results can only be indicative of possible performance improvement. By striking contrast, RUTIN, combining, in somewhat different fashion, the original query with pseudo-relevance feedback and only the web-generated terms that were suggested to the user in CF1, and using InQuery weighting, performs markedly worse than either RUTGBL or RUTGBF3, as well as RUTBE. Because the runs are not strictly comparable, we do not report any significance figures for differences in performance.

**Table 8. Comparison of Combination Runs with Baseline and Baseline Plus Pseudo-RF**

| Run name | R Precision | | Precision at 10 | | Mean Average Precision | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| **RUTBE** | 0.334 | 0.206 | 0.530 | 0.367 | 0.302 | 0.230 |
| **RUTIN** | 0.243 | 0.183 | 0.400 | 0.311 | 0.195 | 0.185 |
| **RUTGBL** | 0.270 | 0.167 | 0.408 | 0.30 | 0.206 | 0.16 |
| **RUTGBF3** | 0.287 | 0.206 | 0.480 | 0.36 | 0.248 | 0.22 |

## 5 Discussion and Conclusions

Summarizing our results, we found partial support for our hypotheses H1-H4, in that expanding our original baseline query derived from topic title and description by adding terms, respectively: derived through clarity analysis of query; derived from the web; derived from both sources; selected by the assessor from both sources; and, generated by the assessor in response to elicitations in a clarification form (in this case, when those terms were given equal weight with the original terms), all significantly increased performance over the baseline run. However, there was no significant improvement of any of these methods of query expansion with respect to the baseline run expanded by straightforward pseudo-relevance feedback, nor was any one of these runs significantly different from any other. Thus, it appears that the best we can say with respect to overall results is that our interventions which require some form of effort from the searcher do improve performance, but not significantly more than automatic methods which require no additional searcher interaction.

The conclusion stated above depends, of course, on the belief that the original query that a searcher will generally provide is more-or-less of the same quality as the query that we generated from the title and description fields of the HARD topics. An alternative view might be that eliciting a query using something like our CF2 in the first instance, would lead to better initial results, which, together with automatic methods, such as pseudo-relevance feedback and our clarity- and web-generated term identification, would lead to more substantial performance improvement. The combination runs lend some support to this idea, but the differences in methods between baseline and combination runs make such a conclusion problematic.

Given that our pseudo-relevance feedback run expanded the query by 100 terms, and that our best performing experimental run, RUTGALL, expanded the query by all of the terms which appeared in both clarification forms, our results might well be explained as simply due to the

well-known phenomena that: longer (reasonable) queries perform better than shorter queries in best-match retrieval; and, combining different sources of evidence leads to increased performance.

All of our results thus suggest that invoking user interaction in order to perform query clarification is unlikely to be cost-effective. That is, if a system elicited richer information problem descriptions in the first instance, which has been shown to be both possible, beneficial and usable (Belkin, et al., 2002), and then enhanced such a description with automatic methods of query expansion such as those that we have described, combining a variety of different sources of evidence, effectiveness of initial information retrieval results could be substantially improved without having to engage the searcher in efforts extraneous to the overall search goal.

## 6 Acknowledgements

We'd like to thank the other members of the Information Interaction Lab who worked on this project: Iliana Chaleva, Li Hui, Narges Kasiri, Ying-Hsang Liu, Marina Malysheva, and Xiangmin Zhang.

## 7 References

Belkin, N.J. (1980) Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, v. 5: p.133-134.

Belkin, N.J., Cool, C, Head, Jeng, J., Kelly, D., Lin, S., Lobash, L., Park, S.Y., Savage−Knepshield, P., Sikora, C. (2000) Relevance Feedback versus Local Context Analysis as Term Suggestion Devices: Rutgers' TREC-8 Interactive Track Experience. In: E.M. Voorhees & D.K. Harman (Eds.) *The Eighth Text REtrieval Conference (TREC 8)* (pp. 565-576). Washington, D.C.: GPO.

Belkin, N. J., Cool, C., Kelly, D., Kim, G., Lee, H.-J., Muresan, G., et al. (2002). Rutgers interactive track at TREC 2002. *Proceedings of the Eleventh Text Retrieval Conference (TREC 2002)*, Gaithersburg, MD.

Belkin, N. J., Kelly, D., Lee, H.-J., Li, Y.-L., Muresan, G., Tang, M.-C., et al. (2003). Rutgers' HARD and web interactive track experiences at TREC 2003. *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, Gaithersburg, MD.

Belkin, N. J., Oddy, R. N., & Brooks, H. (1982). Ask for information retrieval part ii. Results of a design study. *Journal of Documentation, 38*(3), 145-164.

Cronen-Townsend, S., Zhou, Y and Croft, W.B. (2002). Predicting query performance. *In Proceedings of the ACM Conference on Research in Information Retrieval (SIGIR)*, Tampere, Finland, August 2002.

Kelly, D., Dollu, V. D., & Fu, X. (2004). University of North Carolina's HARD track experiments at TREC 2004. *Proceedings of the Thirteenth Text Retrieval Conference (TREC 2004)*, Gaithersburg, MD.

Kelly, D., Dollu, V. D., & Fu, X. (2005, August 15-19). The loquacious user: A document-independent source of terms for query expansion. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, Salvador, Brazil.

Koenemann, J., and Belkin, N. J. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. In Proceedings of the Human Factors in Computing Systems Conference (CHI'96). ACM Press, New York, 1996.

Kwok, K.L., Grunfeld, L., Sun, H.L., Deng, P. and Dinstl, N. (2004). TREC2004 Robust Track Experiments using PIRCS. *In D. K. Harman, editor, Proceedings of the Twelve Text Retrieval Conference, NIST Special Publication,* 2003.

Roussinov, D., and Zhao, L. (2003). Automatic Discovery of Similarity Relationships through Web Mining, *Decision Support Systems,* 35, 2003, pp. 149-166.

Roussinov, D., Zhao, L., and Fan, W. Mining Context Specific Similarity Relationships Using The World Wide Web. *In proceedings of 2005 Conference on Human Language Technologies.*