When Less is More: Relevance Feedback Falls Short and Term Expansion Succeeds at HARD 2005

Fernando Diaz and James Allan
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
[fdiaz,allan]@cs.umass.edu

ABSTRACT

We used clarification forms to study passage term feedback. When compared against pseudo-relevance feedback with an extremely large external corpus, we found that passage feedback resulted in a reduction in performance while term feedback significantly improved recall.

1. OVERVIEW

UMass tested several new techniques in the HARD track this year. First, we developed a new baseline pseudo-relevance feedback technique based on expanding from a large, external corpus [6]. Our results indicate that externally expanding a query can result in improvements over passage feedback. Second, we present a new feedback technique which incorporates both relevance and non-relevance information. Our results indicate that this regularization-based technique can improve performance when used in conjunction with traditional feedback techniques. Third, we present two successful term feedback techniques. Our most successful technique exploits a structured query language in order to significantly improve recall.

2. BASELINE ALGORITHMS

We used the Indri retrieval engine for retrieval experiments [7]. We did not submit any runs which did not perform pseudo-relevance feedback. We experimented with a mixture of relevance models for our baseline ranking algorithm (also introduced in our Robust runs this year) [6]. Readers should consult the refered work for a more thorough description of parameters. Mixture parameters were set to $P(aquaint) = 1$ for MASSbaseTRM3 and $P(bignews) = 1$ for MASSbaseTEE3. No dependence models were used in our baselines.

3. CLARIFICATION FORMS

Our clarification form consisted of several pages of dialog with the searcher. This dialog followed three phases. In the first phase, the searcher was presented with passages from which to judge document relevance. The second phase consisted of term-based feedback; searchers were asked to judge the expected frequency of terms in relevant documents.

3.1 Passage Presentation

In previous years, we found that passages acted as suitable surrogates for documents when being judged for relevance [1]. We divided each of the top 5 documents in the MASSbaseEE3 run into 150-word, half-overlapping passages. We then ranked all of the passages according to query likelihood. The top-ranked passage was considered the document surrogate. Searchers were asked to judge documents as "definitely relevant", "probably relevant", "probably not relevant", "definitely not relevant", and "can't tell". An example page from the passage feedback phase is shown in Figure 1.

3.2 Term Presentation

We conducted an informal study to determine which structured operators were most helpful during query reformulation. The results of this study indicated that one successful technique was to augment the original query with several concept structures. For example, if our original query were "Iraq-Iran War", the reformulated query would include a node representing concepts such as "cities in Iraq", "cities in Iran", "politicians in Iraq", and "politicians in Iran".

We presented 15 terminological feedback pages. Each feedback page in our form started by requesting a judgment for some query concept extracted from the query or—if candidates from the query were exhausted—from the initial retrieval. Terms taken from the initial retrieval were filtered to not include named entities, single letters, or numbers. Because we only presented 15 terms, we ranked all candidates according to their Clarity [2]. These 15 terms represented the core "concepts" of the query.

In addition to the core terms describing the concept, we were interested in presenting terms for clarifying the concepts themselves. For example, if one of the extracted concepts was "salsa", we would like to have the user disambiguate which sense of "salsa" was intended. One method of presenting alternative senses is to use related terms. We accomplished this by first searching WordNet for synsets containing the concept term [5]. We selected the first term from each of the synsets. If the concept word was not found in WordNet or if there were fewer than five synsets detected, we issued the concept word as a query and padded out the related term list with the most frequent terms in this retrieval.

Figure 2 depicts an example term feedback page. Searchers were asked to indicate the expected frequency of terms in relevant documents. Since our core retrieval algorithm is based on term frequencies, we felt requesting this information would be more helpful than asking searchers to judge the more ambiguous concept of "term relevance".

4. PASSAGE FEEDBACK

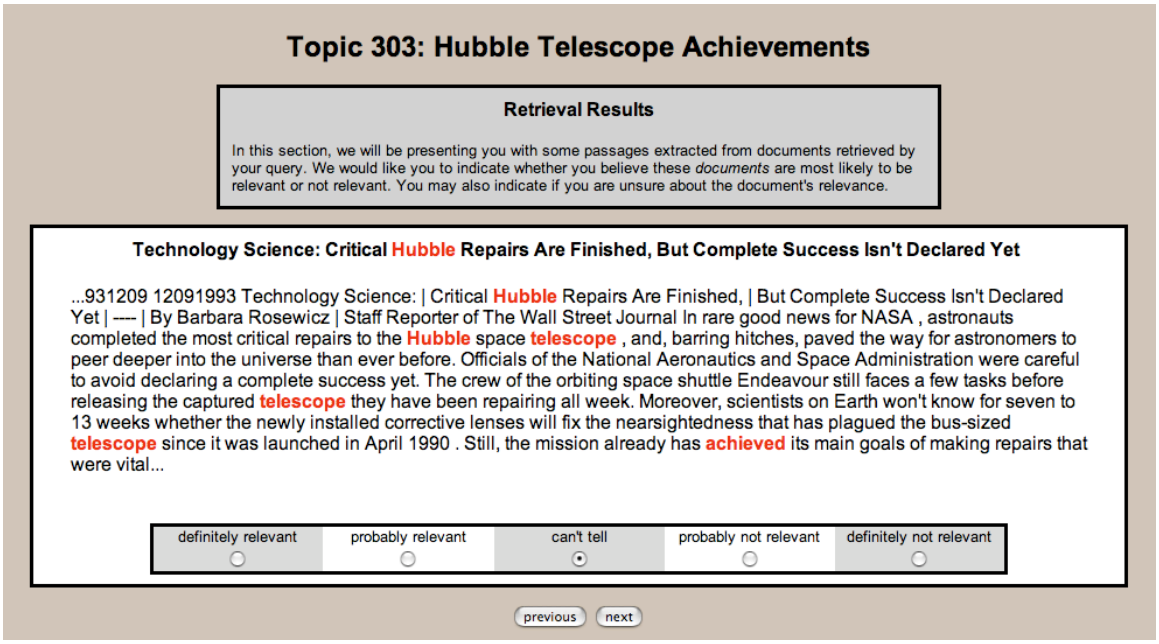

Figure 1: Passage Feedback Interface.

4.1 Passage Relevance Model

Our baseline relevance feedback method builds a relevance model for each set of relevant documents,

$$P(w|\theta_R) \quad = \quad \frac{1}{|R|} \sum_{D \in R} P(w|\theta_D) \qquad (1)$$

After ranking terms in decreasing order of probability, we use the top terms and weights combined with the *original query*. We then perform retrieval with this expanded model. This run is labeled MASSpassRM3.

4.2 Regularizing the Retrieval Scores

In previous work, we found that a second pass of *regularizing* retrieval scores often significantly improves ranking [4]. In order to use this framework, we first normalize retrieval scores using shift and scale normalization. Then, if any judged documents occur in the second retrieval, we replace the retrieval score with 1 for documents marked relevant and 0 for documents marked non-relevant. This run is labeled MASSpassRM3R.

5. TERMINOLOGICAL FEEDBACK

5.1 Regularizing Term Weights

We adopt the regularization framework used in Section 4.2 to regularize the expanded term weights. In this section, we will sketch the notions of term graphs and term weights.

A term graph contains one node for each term in the set of expansion terms. In our case, we used the set of terms in the EE3 model. The edges in the graph are meant to represent the relatedness of two terms. We use a language model approach to measure the relatedness. For each term, we construct a language model based on terms frequently co-occurring with the candidate term in some fixed window of terms. Previous work has found bigram-based distributional similarity to be a compelling method for quantifying term relationships [3]. Our method relaxes the proximity to include words within some window. We accomplished this by retrieving fixed-length passages for each term. We then built a language model out of the weighted combination of maximum likelihood passage models,

$$P(w|\theta_q) \quad = \quad \sum_{T \in R} P(w|\theta_T) \frac{P(q|\theta_T)}{\mathcal{Z}} \qquad (2)$$

where $\mathcal{Z}$ is a normalizer over all retrieved passages, $T \in R$. We use the multinomial diffusion kernel to compare language models, in turn defining our edge weights [4]. Figure 3 shows an example graph.

Once this graph is constructed, we can regularize the term weights used in MASSbaseEE3 following the feedback the method described in Section 4.2.

5.2 Building Structured Queries

As mentioned in Section 3.2, we conducted an informal study to determine useful patterns in structured querying. One important technique we found was the use of concepts. Instead of querying with a set of terms, we would, for each concept in the query, expand the term into conceptually related terms. We developed the concept-feedback template presented in Figure 4.

6. FREE TEXT FEEDBACK

We also provided the searcher with an opportunity to manually submit additional query terms in a free text box. No searchers managed to reach this phase of the clarification process so we exclude it from our results.

7. TRAINING

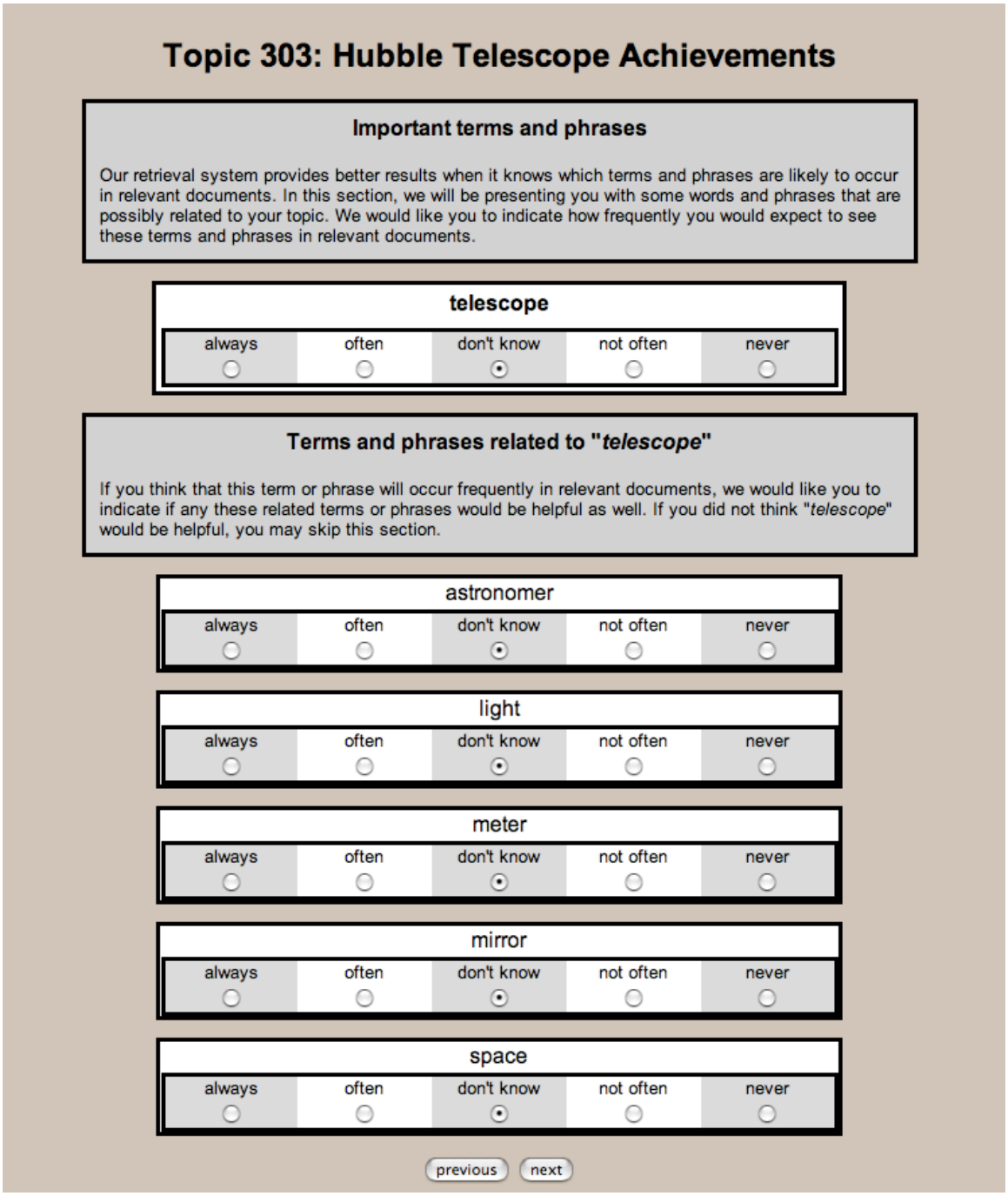

Figure 2: Term Feedback Interface.

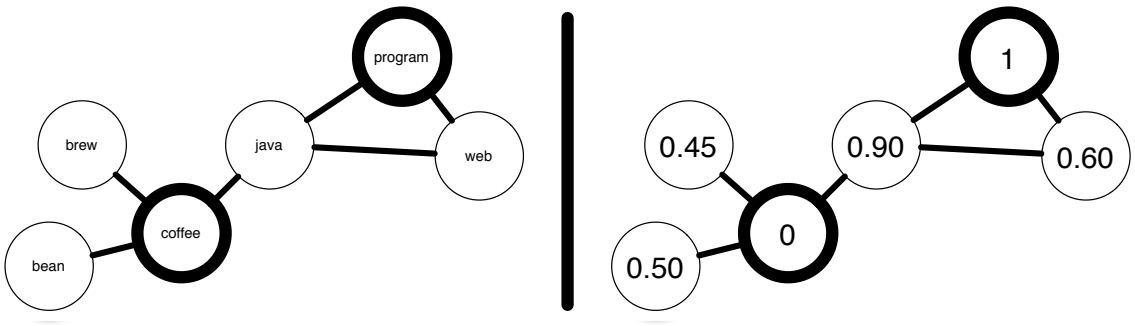

Figure 3: Term Regularization. The figure on the left represents co-occurrence information of terms in the expansion model. Highlighted nodes represent terms presented to the user for feedback. The figure on the right represents the pre-regularization term weights. The weight of highlighted nodes depends on the user feedback (0 for non-relevant terms; 1 for relevant terms). The weight of all other terms in the graph is equal to the expansion model weight.

```
#weight( LAMBDA1 #combine(EE)
         (1-LAMBDA1) #combine( #wsyn( LAMBDA2 #1(PHRASE1) (1-LAMBDA2) #syn(RELATEDTERM11 RELATEDTERM12...))
                               #wsyn( LAMBDA2 #1(PHRASE2) (1-LAMBDA2) #syn(RELATEDTERM21 RELATEDTERM22...))
                               ...
                             )
       )
```

Figure 4: Structured Query Template. The EE model is generated from the external corpus. The phrases are header terms in the clarification form. The related terms are footer terms in the clarification form. Weights for λ_1 and λ_2 are learned in the training phase.

We used the Robust 2004 topics and corpus for training parameters. We only considered the title field for our training experiments. We excluded topics used in Robust 2005 evaluation. We simulated the HARD user study by asking 10 graduate students in our department to complete clarification forms for 10 topics in three minutes. This resulted in a total of 100 topics to use for training feedback algorithm parameters.

7.1 Baseline

Our baseline pseudo-relevance feedback retrieval used 10 feedback documents to generate the relevance model, 25 terms from the relevance model to interpolate with the original query, and a weight of 0.20 on the original query. When using the external corpus, the retrieval used 25 feedback documents to generate the relevance model, 50 terms from the relevance model to interpolate with the original query, and a weight of 0.10 on the original query. The values of these trained parameters indicate that we should have more confidence in the relevance model built from the external corpus.

7.2 Passage Feedback

We trained our passage Feedback algorithms using *document* judgments from the Robust 2004 topic set and corpus. This allowed us to train on the complete set of 200 topics (after having excluded the topics used in Robust 2005). Although our clarification forms resulted in passage feedback, we found this training technique successful in previous years [1]. Our feedback algorithm used only relevant documents to generate the relevance model, 50 terms from the relevance model to interpolate with the original query, and a weight of 0.10 on the original query. Our regularization used a 5-nearest neighbor graph, $t = 0.70$, and $\alpha = 0.20$.

7.3 Term Feedback

We trained both of our term feedback algorithms using the in-house clarification form data. In both cases, only terms labeled "always" and "never" were used for feedback.

Our term regularization model had two components to train: term language models and term regularization. When build the per-term language models, we used the top 100 50-word passage retrieved. Each graph consisted of the top 100 terms from the relevance model. Our regularization used a 75-nearest neighbor graph, $t = 0.50$, and $\alpha = 0.50$.

Our structured term feedback model had two parameters. The original query was given the most weight with $\lambda_1 = 0.95$. Within the concept node, we found that more weight was placed on the related terms with $\lambda_2 = 0.30$. We performed a post-processing step where all terms labeled "never" were removed from the structured query.

8. RESULTS

We submitted two baseline runs and four post-clarification runs. All runs used the title field only. The results are presented in Table 1. We present standard retrieval measures including the official R-Precision measure.

8.1 Baseline Runs

Our baselines performed similarly to our Robust track runs. Expanding the query using only the Aquaint corpus

	MASSbaseTRM3	MASSbaseTEE3	MASSpsgRM3	MASSpsgRM3R	MASStrmR	MASStrmS
Retrieved:	50000	50000	50000	50000	50000	50000
Relevant:	6561	6561	6561	6561	6561	6561
Relret:	4115	4357	4241	4241	4877*	**4909***†
IntPrec@0.00	0.6302	0.7088	0.7257*	0.7013	**0.7603***	0.7584*
IntPrec@0.10	0.4548	0.5631	0.5049	0.5214	0.5798*	**0.5925***
IntPrec@0.20	0.3997	0.4897	0.4336	0.4478	0.4917*	**0.5270***
IntPrec@0.30	0.3547	0.4392	0.3791	0.3861	0.4222*	**0.4711***
IntPrec@0.40	0.2944	0.3783	0.3221	0.3310	0.3643*	**0.3988***
IntPrec@0.50	0.2474	0.3076	0.2718	0.2808	0.3001*	**0.3202***
IntPrec@0.60	0.2011	0.2420	0.2146	0.2253	0.2379	**0.2605***
IntPrec@0.70	0.1625	0.1897	0.1659	0.1734	0.1902	**0.2042***
IntPrec@0.80	0.1111	0.1256	0.1091	0.1134	0.1250	**0.1419***
IntPrec@0.90	0.0524	0.0639	0.0497	0.0534	0.0659	**0.0662***
IntPrec@1.00	0.0068	0.0037	0.0031	0.0029	0.0057	**0.0070**
map	0.2445	0.3043	0.2688	0.2766*	0.3019*	**0.3223***
P@5	0.4360	0.5600	0.5160*	0.5200*	**0.5640***	0.5600*
P@10	0.4300	0.5300	0.4780	0.4880	0.5320*	**0.5600***
P@15	0.4013	0.5253	0.4707*	0.4933*	0.5093*	**0.5333***
P@20	0.3970	0.5050	0.4610*	0.4740*	0.4940*	**0.5110***
P@30	0.3780	0.4767	0.4347	0.4453*	0.4647*	**0.4853***
P@100	0.2848	0.3580	0.3256*	0.3340*	0.3566*	**0.3732***
P@200	0.2251	0.2605	0.2454	0.2480*	0.2731*	**0.2830***
P@500	0.1363	0.1498	0.1430	0.1443	0.1614*	**0.1651***
P@1000	0.0823	0.0871	0.0848	0.0848	0.0975*	**0.0982***†
rprec	0.2660	0.3291	0.3024*	0.3082*	0.3353*	**0.3547***

Table 1: Results. Comparisons between official baseline and modified retrieval runs. The first two columns represent local pseudo-relevance feedback (MASSbaseTRM3) and external pseudo-relevance feedback (MASSbaseTEE3). The second two columns represent passage feedback (MASSpsgRM3) and regularized passage feedback (MASSpsgRM3R). The final two columns represent feedback using term regularization (MASStrmR) and structured query construction (MASStrmS). Bold numbers represent the best performance among our systems. Statistical improvements were computed with respect to both baselines. We used a Wilcoxon test and indicate instances where $p < 0.05$. A superscript $*$ indicates improvement over MASSbaseTRM3 and a superscript † indicates improvement over MASSbaseTEE3.

resulted in an R-Precision much lower than external expansion.

8.2 Passage Feedback

The feedback runs perform better than our pseudo-relevance feedback baseline (MASSbaseRM3). Note that candidate feedback documents were taken from the external corpus. As expected, regularization boosted the relevance feedback performance across almost all measures.

One result for passage feedback is surprising. Feedback does not outperform external expansion (MASSbaseEE3). This result seems to indicate that true relevance feedback underperforms pseudo-relevance feedback when we gather an external corpus of sufficient size and quality. We are currently conducting experiments to determine the precise situations when external expansion is better than true feedback.

8.3 Term Feedback

Our best runs incorporated term feedback. This is surprising since our previous work indicated that term feedback often did not improve retrieval; only free text reformulation provided performance gains. Because we have adjusted both our term selection and term incorporation scheme from previous years, it is difficult to determine which factor is likely to explain the improvement.

Several improvements over the MASSbaseEE3 baseline should be pointed out. First, although both term feedback methods improved the official R-Precision metric, only structured feedback improved mean average precision. Second, the only significant improvement over our external expansion baseline was in the number of relevant documents retrieved. There are several reasons for this. Term presentation always occurred after document presentation. Therefore, we tended to have far fewer term judgments. In several cases, searchers did not even get to the term feedback pages. Since additional term information existed for fewer topics, then, it was difficult to measure the significance of this improvement.

That the number of relevant documents significantly rose is contrary to previous year's results where additional terms only improved high precision measures and sometimes reduced recall. We believed that the searchers were adding terms recognized as discriminative in the top-presented passages on our form or such terms seen on other sites' forms in the study. The result would be to construct a high-precision query for retrieving those documents already included on the clarification form(s). We constrained searchers this year to a small set of terms unlikely to be prone to such over-fitting. Because of it's query-directed nature and the named-entity filtering, the candidate set of terms was much more general. Combined with our structured query, we could carefully expand these general concepts to retrieve a larger mass of relevant documents.

9. CONCLUSIONS

Our experiments this year provided some compelling results. First, we demonstrated that passage-level feedback sometimes under-performs pseudo-relevance feedback using an external corpus. This is interesting because it means that there are cases where document feedback might be less useful than exploiting some larger or higher quality corpus. Second, we showed that term feedback can be successfully used to improve recall tasks whereas previous results demonstrated only precision gains.

10. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval, in part by the Defense Advanced Research Projects Agency (DARPA), and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the authors and do not necessarily reflect those of the sponsor.

11. REFERENCES

[1] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. Umass at trec 2004: Novelty and hard. In *Online Proceedings of 2004 Text REtrieval Conference*, 2004.

[2] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2002.

[3] I. Dagan, L. Lee, and F. Pereira. Similarity-based models of cooccurrence probabilities. *Machine Learning*, 34(1-3):43–69, 1999.

[4] F. Diaz. Regularizing ad hoc retrieval scores. In *CIKM 2005: Proceedings of the fourteenth international conference on Information and knowledge management*. ACM Press, 2005.

[5] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[6] D. Metzler, F. Diaz, T. Strohman, and W. B. Croft. Umass at robust 2005: Using mixtures of relevance models for query expansion. In *The Twelth Text REtrieval Conference (TREC 2005) Notebook*, 2005.

[7] T. Strohman, D. Metzler, H. Turtle, and W. B. Croft. Indri: A language model-based serach engine for complex queries. In *Proceedings of the International Conference on Intelligence Analysis*, 2004.