IIT at TREC-8: Improving Baseline Precision

M. Catherine McCabe
Advanced Analytic Tools
Washington, DC
catherm@ir.iit.edu

David O. Holmes
NCR Corporation
Rockville, MD
david.holmes@washingtondc.ncr.com

Kenneth L. Alford
US Army
Springfield, VA
ken4sher@erols.com

Abdur Chowdhury
IIT Research Institute
Rockville, MD
abdur@ir.iit.edu

David A. Grossman
Illinois Institute of Technology
Chicago, IL
dagr@ir.iit.edu

Ophir Frieder
Illinois Institute of Technology
Chicago, IL
ophir@ir.iit.edu

Abstract

In TREC-8, we participated in the automatic and manual tracks for category A as well as the small web track. This year, we focussed on improving our baseline and then introduced some experimental improvements. Our automatic runs used relevance feedback with a high-precision first pass to select terms and then a high-recall final pass. For manual runs, we used predefined concept lists focussing on phrases and proper nouns in the query. In the small web-track, we submitted one content-only run and two link-plus-content runs. We continued to use the relational model with unchanged SQL for retrieval. Our results show some promise for the use of automatic concepts, expansion within concepts and a high-precision first pass for relevance feedback.

1. Introduction

Our work for TREC-8 is a continuation of the work started in TREC-3 when we implemented an information retrieval system as an application of a relational database management system (RDBMS). We used unchanged Structured Query Language (SQL) to implement vector-space relevance ranking [Grossman95, Grossman96]. TREC-4 work demonstrated the relational implementation on category A data and introduced the concepts-list approach in the manual runs. In TREC-5, we implemented relevance feedback and entered the Spanish, Chinese and Confusion tracks. For TREC-6, we expanded our relevance feedback methodology to include the lnc-ltc term weights [Singhal96]. During TREC-6, we explored the assumption that certain infrequently occurring terms with high collection weights may actually be artificially inflating the query-to-document relevance ranking scores. We continued that work in TREC-7 with expanded stop lists and term thresholding. In addition, with TREC-7 we combined information extraction (IE) techniques with information retrieval through the use of a relevance feedback filter based on IE. During each of those years, our system performed well, but we noted that our baseline results were below those of other teams using similar retrieval strategies. So this year, we focused first on improving our baseline and then on experimentation with automated concepts and various expansion techniques, including a high-precision first-pass relevance feedback technique.

We began entering the manual track in TREC-4. This effort has focussed on structuring queries via concepts and manual relevance feedback while spending less than one half hour on each query. In TREC-5, we experimented with the use of manually assigned term weights. For TREC-6, we used inexact term matching and an automatically generated thesaurus based on term co-occurrence. In TREC-7, our manual run focused on using phrases and proper nouns within the concepts. In addition, a more detailed iterative process was introduced. These manual techniques landed us among the top participants in manual track for TREC-7. This year, we continued the successful techniques and worked to ensure that we added key proper nouns and phrases for each concept in the query

We participated in the small web track introduced this year. Our relational platform proved to be quite flexible and was able to index the web documents with minor changes to the pre-processor (parser.) Our baseline (content-only) run used the straightforward vector space model with Singhal's pivoted cosine normalization [Singhal96]. Our experimental (link-plus-content) runs used link information to weight and rerank documents retrieved.

2. Prior Work in Relational IR

The implementation of an Information Retrieval (IR) system using the relational model hinges on the use of a relation (table) to model an inverted index, which is the central data structure in traditional IR systems. The traditional inverted index stores each unique term or phrase from the collection and a list of all the documents containing each term/phrase. The inverted index can also include frequency, offset, or other desired information. In the relational approach, this index is normalized and stored in a table. Queries using standard structure query language (SQL) are used to find and rank all documents containing the query terms. Full details of the implementation can be found in Grossman97 and Lundquist97. One benefit to using the relational model for IR is the ability to exploit parallel processing via the DBMS. All commercial DBMS systems offer a parallel version. For TREC-8, our manual runs used Windows NT versions of NCR/Teradata and Sybase/Adaptive Server Enterprise on Pentium SMP servers. This year's ad hoc and small web track submissions used Oracle on SUN Solaris machines.

3. Implementation Details

In this section we first discuss the baseline improvements made to our system and then we present our work in each track – automatic ad hoc, manual ad hoc and the small web track.

3.1. Improving the Baseline

In this year's work, we focused on the fundamentals and conducted many comparisons with the best systems from last year's TREC. The baseline title+description runs for the top three performers at TREC-7 were OKAPI 0.233, ATT 0.218 and UMASS 0.20. Our own baseline for TREC-7 queries was 0.17. We looked for system differences to explain this lower performance. We began by examining the difference that the retrieval strategy makes. We implemented the same probabilistic retrieval strategy as given in [Robertson98]. We found that average precision recall did not differ significantly from previous runs using vector space strategies. We analyzed the result sets and found that they had very high overlap (relevant and nonrelevant) and very similar rankings. We concluded that the different retrieval strategies (when based on tf*idf) do not account for the differences in average precision recall.

We next considered the impact of token selection. Various stemming, phrases, and thesaurus techniques impact the tokens that represent the documents and the queries. We noted that the GSL file was instrumental in the OKAPI systems token selection -- conflating acronyms with their terms, American and British term variants, as well as many synonym groups. The GSL file only affected a few TREC-7 queries, but it had a large positive impact on almost all that it affected. In addition, the leading systems all used stemming approaches, while we did not. Phrase usage varied across the systems and was reported to result in a .1 to .2 improvement over terms, which is consistent with our own phrases. Stop lists also varied across the systems but it was unclear that this impacted precision/recall. We experimented in all of these areas, and found the keys improving our baseline were the ‘stemming’ and our title-phrase generation. We used the kstem+Porter equivalence groups developed at UMASS to add term variants to the query [Allan98]. This ‘stemming’ was quite effective and landed us at 0.196 average precision recall for title+description.

Our phrase generation technique uses every pair-wise combination of title terms. These new phrases helped (although not by much) most TREC-7 queries and did not cause serious degradation on any query. So we kept the technique for our TREC-8 runs. Finally, we had reached 0.20 and decided this was close enough (matching the third best) and moved on to query expansion.

3.2 Automatic Runs: High Precision Relevance Feedback with Automated Concepts

To ensure the top documents used for selecting expansion terms were relevant, we implemented a high-precision filter. This filter set up a concept for each title query word, used the Porter/k-stem algorithm to expand terms in each concept, and then required a document to contain at least one term from each concept. For example, the query 401, “foreign minorities, germany”, results in three “concepts” created: 1) foreign, foreigner, foreigners; 2) minority, minorities 3) german, germany. The high precision first pass requires at least one word from each concept to be present for a document to qualify. Essentially, this is a logical AND of several OR groups. Ranking was achieved with the usual vector space similarity measure.

To select terms from these top documents, we used a modified Rocchio approach with the additional filter of requiring the term to occur in at least 2 of the top documents. We experimented with the number of top documents and number of terms to use and found that 10 terms from documents was best (see Tables 1 and 2.)

We note that similar work has been done earlier (most notably [Mitra98]) but our specific variations (automatic title concepts expanded with k-stems and N > 1) are new and effective.

Test	Average Precision
Using top 1 doc	.1609
Using top 2 docs	.2287
Using top 10 docs	.2359

Table 1. Calibration of Relevance Feedback using 10 Terms (TREC-7)

Test	Average Precision
No Feedback	.1966
Add 10 terms	.2359
Add 20 terms	.2065
Add 30 terms	.2057
Add 40 terms	.2057
Add 50 terms	.2100

Table 2. Calibration of High-Precision Relevance Feedback using 10 Documents (TREC-7)

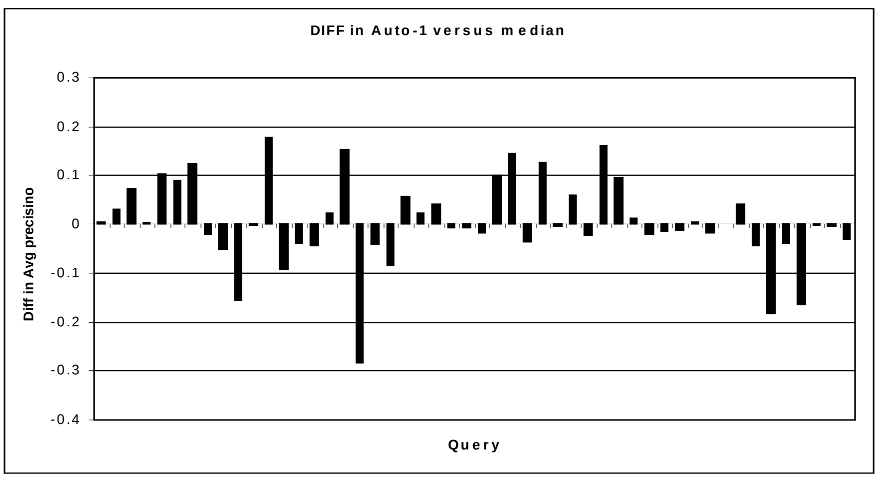

Figure 1. IIT Automatic Run-1 Difference from the Median

When we ran the second pass, we loosened the restriction of requiring at least one word from ALL title concepts to requiring at least one word from any ONE of the concepts. At this point our average precision recall was up to 0.2359. Finally, we reranked our resulting set of documents by the percentage of query terms found in the document. This reranking gained a small improvement, bringing our final TREC-7 run to 0.2454.

3.3. Manual Run

We spent approximately one-half hour formulating each query for our manual runs, using an iterative, interactive approach. The searcher used manual relevance feedback and general knowledge to identify new query words.

3.3.1. Manual Run Implementation Details

Consistent with previous years, our manual effort separated each query into a set of concepts -- search, scoring and negation. Search concepts are used in a vector space retrieval as a first pass. Terms from scoring concepts are then added to the document vectors and the documents are reranked. Finally, the negation concept 'disqualifies' a document from the result set. For TREC8, we used the negation concept more frequently than ever before—in 34 of the topics. Negation concepts included 147 phrases and 63 single words. Search concepts included 155 words and 498 phrases. The remaining tokens comprised the scoring-only concepts. The negative concept technique eliminated many irrelevant documents from our results. For example, on Topic 447 (*Stirling Engine*) we achieved 1.0 average precision recall by eliminated documents about *Stirling University* and people with the surname of *Stirling.*

	Search Concepts	Scoring Concepts	Negation Concepts	TOTAL
Terms	155	352	63	570
Phrases	498	567	147	1,212
Total	653	919	210	1,782

Table 3. Use of Concepts in Manual Track

As in our TREC-7 work, we emphasized phrases and proper nouns in the IIT manual ad hoc queries. For TREC-8, we used 1,782 search tokens including 1,212 phrases. Half of the phrases were proper nouns and the remaining were mostly common noun phrases. Of the 570 single words, 508 were either common or proper nouns. In other words, 96.5% of all search tokens were either phrases or single word nouns.

3.3.2. Manual Run Analysis

The average precision for our manual run was officially scored at 0.4104. We were at or above the median on 38 of 50 queries. When we were below the median, it was by a small margin and when we were above it was on average, by a much larger margin. We conducted failure analysis to determine why some queries performed poorly and why some did very well. This year, we spent more time reading the documents retrieved than any other year and believed most documents in the results sets to be relevant to the query. During our query development phase, the analyst tagged documents as relevant, doubtful, or non-relevant. We compared our list to the official results and found numerous differences (summarized in Table 4). Document relevance is subjective, of course, and subject to interpretation, but several of the differences in evaluation were difficult to reconcile. For example, Topic 423 asked for any references to Mirjana Markovic, the wife of Slobodon Milosevic – "*Any mention of the Serbian president's wife is relevant*". We found document FT942-3554 and FBIS3-2, both of which mention her by name and yet were judged non-relevant (see Figure 2).

NIST Relevance Assessment	IIT Relevance Assessment	Number of Documents
Relevant	Relevant	895
Relevant	Doubtful	188
Relevant	Non-Relevant	99
Non-Relevant	Relevant	428
Non-Relevant	Doubtful	356
Non-Relevant	Non-Relevant	1033

Table 4. Comparison of Relevance Judgments

Clearly such inaccuracies in relevance assessment have an impact on average precision recall. The average precision recall for our manual run increases to around .4800 when we use our relevance assessments.

<DOCNO> FT942-13554 </DOCNO> *taken from the text:*	<DOCNO> FBIS3-2 </DOCNO> *taken from the text:*
". . .Of special interest in Duga is the diary of **Mrs Mirjana Markovic, the wife of Mr Milosevic.** Her musings on the nature of life, spring-time in Belgrade often sound the death knell for the political rivals of her husband or herald an imminent Machiavellian manoeuvre by the Serbian President. The diary of Mrs Markovic is then reprinted in Politika, the oldest and most influential Serbian daily. . ."	". . . Independent biweekly that carries political and social commentary as well as articles focusing on popular culture. Regularly carries a column of political commentary written by **Mirjana Markovic--Milosevic's wife**—that often criticizes the Serbian nationalist cause. . . ."

Figure 2. Sample Judgments for Topic 423

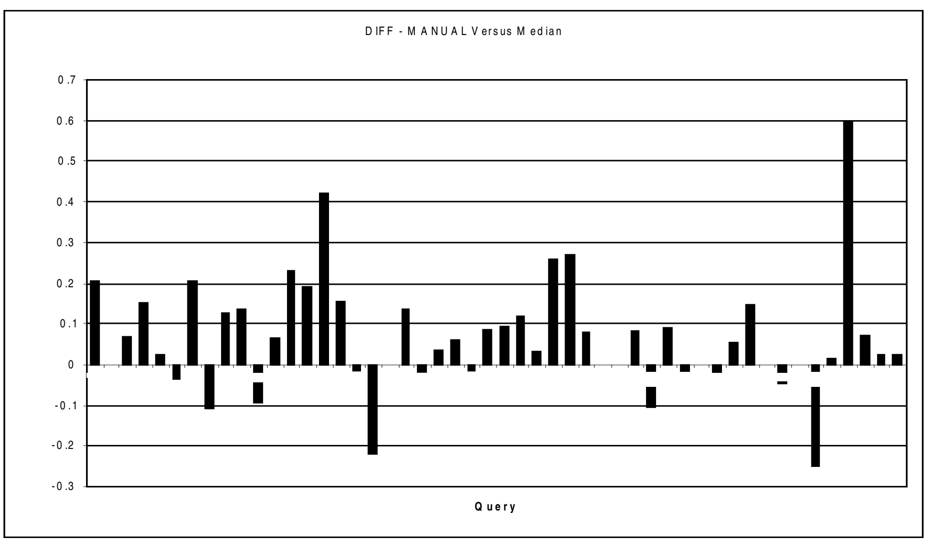

Figure 3. IIT Manual Run Difference from the Median

3.4. Small Web Track

This year we entered the new small web track. We used our baseline information retrieval system with only minor changes to the preprocessor. In this section, we describe our techniques, results and analysis.

3.4.1. Small Web Track Implementation Details

Our Content-Only run (iit99wt1) simply used our baseline relational IR system to process the small web collection, using the title+description queries. The Link-Plus-Content runs (iit99wt2 and iit99wt3) began with the document sets retrieved during the Content-Only runs and then incorporated link data and reranked the results. Many of our initial efforts to incorporate links to or from other web pages resulted in reduced average precision values when measured against the TREC-7 benchmark data. We observed that the highest concentration of relevant retrieved documents occurred near the beginning of the documents retrieved for each topic; therefore, there was little or no need to reorder those high-ranking documents. We then retrieved documents beyond the original 1000 documents per query. We sought to use web links to identify and add documents to the result set. The approach we used was similar to the *root set* proposed in [Kleinberg97]. The top *x* documents (50 for Run-1, iit99wt2, and 100 for Run-2, iit99wt3) were included in the root set. The root set was then expanded so that links to and from those documents were added to the set of retrieved documents *if* they were already present in the set of all documents retrieved for a specific topic. In order to keep the result set within the maximum 1000 documents per topic, the lowest ranking documents from the original Content-Only run were removed from the result set. New documents were weighted and added to the retrieved documents set in such a manner that their original rankings were retained within the new result set.

Run Description	Relevant Retrieved	Average Precision
Content-Only	4480	0.2817
Link-Plus-Content	4523	0.2861

Table 5. IIT Small Web TREC-7 Benchmarks

3.4.2. Small Web Track Results

A comparison of results from our small web track runs is shown in Table 6. Our Content-Only run (iit99wt1) scored below the median on 27 of the 50 topics. We were neither the best nor the worst on any topic. When compared again against the median, our performance for Run-2 (iit99wt2, Link-Plus-Content) was greatly improved over the Content-Only run (iit99wt1). We received the best average precision score on three topics (419, 423, and 435) and were equal or above the median on 34 of the 50 queries. Since our average precision remained the same as the Content-Only run (at 0.2265), the relative improvement over the median is due to other teams degrading in their link-based runs.

Run Description	Run Identifier	Average Precision	Judged Relevant	Relevant Retrieved
Content-Only	iit99wt1	.2265	2279	1575
Link-Plus-Content (Run 1)	iit99wt2	.2265	2279	1572
Link-Plus-Content (Run 2)	iit99wt3	.2264	2279	1568

Table 6. IIT Small Web TREC-8 Results

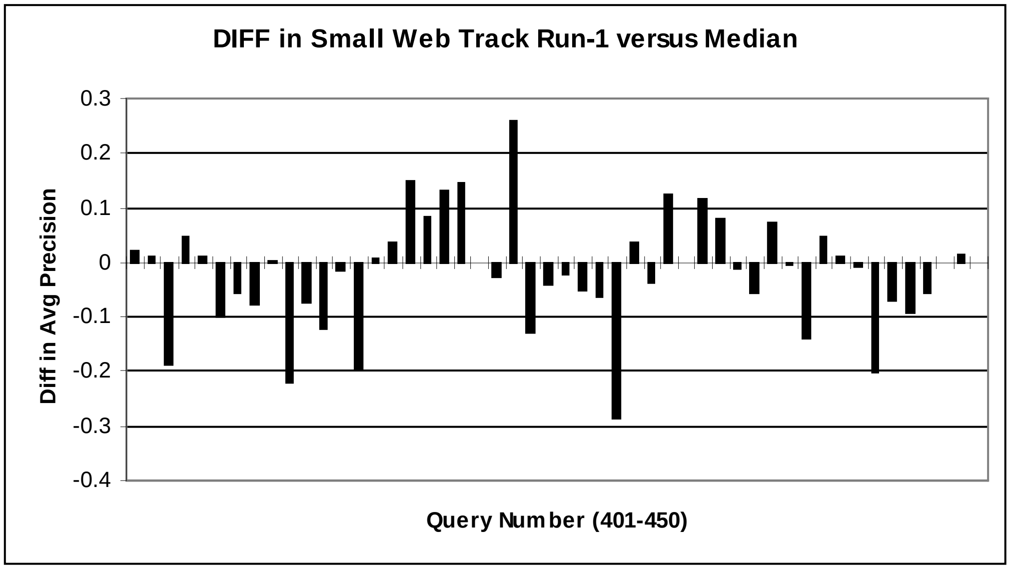

Figure 4. IIT Small Web Track Run-1 (Content-Only) Difference from the Median

3.4.3. Small Web Track Analysis

Incorporating link information is a challenging problem. As numerous studies have noted, all web links are not of equal value [Spertus97, Kleinberg97]. We have not yet found an effective way to automatically evaluate and discriminate between the numerous types of links that exist within web-based documents. Our excellent performance on query 423 can be attributed to the underlying retrieval engine and not to any specific techniques for web documents. We did well on it for the concept only run as well as for the link-based runs. The same can be said for our poor performance on query 403 and 429. An interesting factor in analyzing web track results is found in the sparseness of the qrels set. The TREC-8 qrels set is only 35 percent as large as the TREC-7 qrels set (2279 vs. 6495) and only 50 percent as large as the ad hoc track. (2279 vs. 4728).

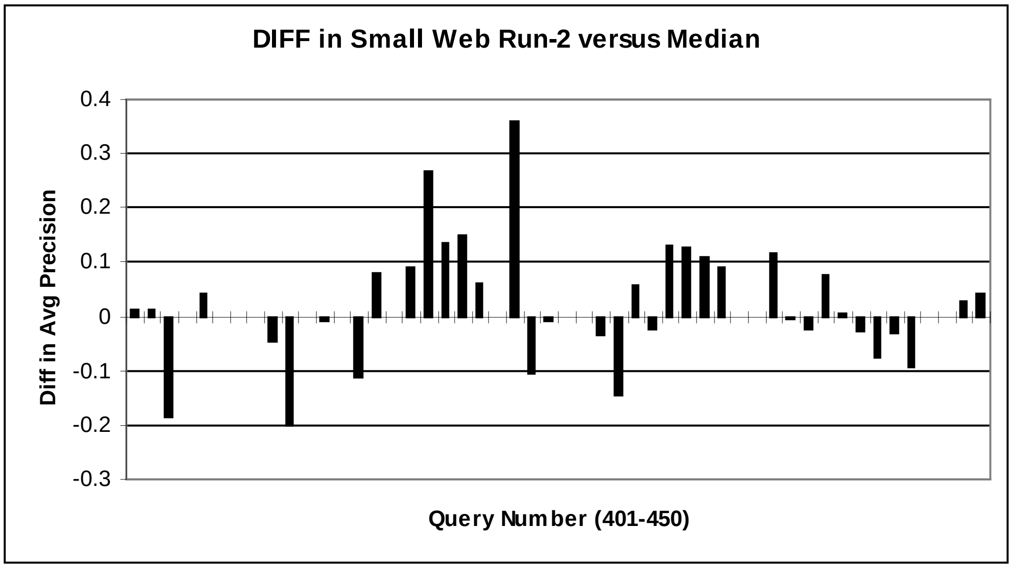

Figure 5. IIT Small Web Track Run-2 (Link-Plus-Content) Difference from the Median

4. Conclusions and Future Work

For TREC-8, we focused on improving our baseline system and then introducing some new feedback techniques. We identified key enhancements to our parser and our feedback engine. We introduced a technique for using k-stem conflation to expand title-term concepts and use this as a filter for high-precision relevance feedback. Our success in the manual track shows that phrases and nouns are important elements in runs with high average precision. Our work in the web track was a good beginning, but our results highlight the fact that there is still much room for improvement. Adjusting content runs based on link information assumes accurate content-only results and link information that can effectively weight and rank those results. Research will continue to improve both elements.

Table 7 summarizes the results of IIT TREC-8 submissions.

	iit99au1 (Tit+Des)	iit99au2 (Tit+Des)	iit99ma1 (Manual)	iit99wt1 (Content)	iit99wt2 (Link-Plus)
TREC-8 Track	Ad Hoc	Ad Hoc	Manual	Sm Web	Sm Web
Avg. Precision	0.2305	0.2041	0.4104	0.2265	0.2265
Precision at 10 Documents	0.4749	0.4343	0.7790	0.4100	0.4100
Documents Judged Relevant	4728	4728	4728	2279	2279
Relevant Retrieved	2688	2207	3106	1575	1572
At or Above Median (Avg. Prec.)	23	-	37	23	34
Below Median (Avg. Prec.)	27	-	13	27	16

Table 7. IIT TREC-8 Results Summary

Our future challenges include: (1) further integration of information extraction in relevance feedback, (2) the need to move beyond proper nouns and experiment with the use of entities as feedback filters, and (3) methods to more effectively evaluate and weight link information. In addition, the automation of the manual techniques used to add high quality phrases into our searches is an area for future work.

Acknowledgments

We wish to thank the director and staff at the Major Shared Resource Center, U.S. Army Research Lab, Aberdeen, MD for their generous support in making the small web track research possible.

References

(Allan98) Allan, J.A., J. Callan, M. Sanderson, J. Xu, and S. Wegmann. "Inquery and TREC-7". *Proceedings of the Seventh Text REtrieval Conference (TREC)*, sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1998.

(Buckley95) Buckley, C. A. Singhal, M. Mitra, and G. Salton, "New Retrieval Approaches Using SMART: TREC-4," *Proceedings of the Fourth Text REtrieval Conference (TREC),* sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1995.

(Fox90) Fox, Christopher. A Stop List for General Text. *SIGIR Forum*, (v. 24, no. 1-2) 1990, p. 19-35.

(Grossman95) Grossman, D., D. Holmes, O. Frieder, M. Nguyen, and C. Kingsbury, "Improving Accuracy and Run-Time Performance for TREC-4," *Proceedings of the Fourth Text REtrieval Conference (TREC)*, sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1995.

(Grossman96) Grossman, D., C. Lundquist, J. Reichert, D. Holmes, and O. Frieder, "Using Relevance Feedback within the Relational Model for TREC-5," *Proceedings of the Fifth Text REtrieval Conference (TREC)*, sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1996.

(Grossmam97) Grossman, D., D. Holmes, O. Frieder, and D. Roberts, " Integrating Structured Data and Text: A Relational Approach," *Journal of the American Society of Information Science*, January 1997.

(Kleinberg97) Kleinberg, Jon M. "Alternative Sources in a Hyperlinked Environment," *IBM Research Report (RJ-10076)*, May 29, 1997.

(Lundquist97) Lundquist, C., D. Grossman, O. Frieder, and D. Holmes, "A Parallel Implementation of Relevance Feedback using the Relational Model," *Proceedings of the World Multiconference on Systemics, Cybernetics, and Informatics*, July 1997.

(Lundquist98) Lundquist, C., D. Holmes D. Grossman O. Frieder. "Expanding relevance feedback in the relational model." *NIST Special Publication 500-240*, pages 489-502, August 1998.

(Mitra98) Mitra, M., A. Singhal, C. Buckley. "Improving Automatic Query Expansion". *Proceedings of the Twenty-First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* ACM SIGIR'98 pages 206-214, 1998.

(Robertson98) Robertson, S.E., S. Walker, M. Beaulieu. "Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track". *Proceedings of the Seventh Text REtrieval Conference (TREC)*, sponsored by the National Institute of Standards and Technology and the Advanced Research Projects Agency, November 1998.

(Singhal96) Singhal, A., C. Buckley, and M. Mitra, "Pivoted Document Length Normalization," *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Ed. Hans-Peter Frei, Donna Harman, Peter Schauble and Ross Wilkinson, August 18-22, 1996.

(Spertus 1997) Spertus, E. "ParaSite: Mining Structural Information on the Web," *HyperProceedings of the Sixth International World Wide Web Conference*. Electronic copy: http://atlanta.cs.nichu.edu.tw/www/PAPER206.html, 1997.