IIT at TREC-8:
Improved Manual Query Processing and
Using Stemming Equivalence Classes as a Basis for Relevance Feedback

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 first ensured that our baseline matched the effectiveness achieved by other teams using the same ranking techniques. We then introduced some experimental improvements. We investigated differences among the top TREC participants from past years and corrected some minor variations in our system. For the automatic runs, we included a baseline run (iit99au1) and an experimental run (iit99au2) that used a concept-based expansion technique. The automatic runs used the required title plus description ('short') query versions. The experimental run 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 with terms from the concept lists combined in different ways. The manual run focused on using 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 with this year's automatic ad hoc system using Oracle and the manual ad hoc using both Teradata and Sybase DBMS. 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. TREC-5 also used the relational approach for the Spanish, Chinese and Confusion tracks. For TREC-6, we expanded our relevance feedback methodology to include the lnc-ltc term weights [Singhal96] as well as feedback term scaling. 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 techniques with information retrieval through the use of a relevance feedback filter based on IE (Information Extraction). During each of those years, our system performed well, but we noted that our baseline results were somewhat below those of other teams using similar retrieval strategies. So this year, we focused first on improving our baseline and then on experimentation with our automated concepts and various expansion techniques, including a high-precision first-pass relevance feedback technique.

Our manual runs have focused on the concept approach to structuring queries. In TREC-4, we assigned the query terms into concept lists and used words obtained from various sources (dictionaries, newspapers, etc.) to expand the query to include other similar terms not found in the topic. In TREC-5, we continued to use the concept lists and experimented with the use of manually assigned weights to the query terms as well as using manual relevance feedback to identify additional terms. For TREC-6, we augmented our prior work with inexact term matching and an automatically generated thesaurus based on term-to-term co-occurrence. In TREC-7, our manual run took a somewhat more structured approach than in years past, with the hope of automating some techniques. In particular, our manual run focused on using phrases and proper nouns to improve precision and recall. A more detailed iterative process was used in which we examined initial results and worked to quickly identify new queries. 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. Our results in this area have been encouraging in that the amount of time we spend on each query is typically under a half-hour. 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 after minor changes were made to the pre-processor (parser.) We submitted three small web track data runs. 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 reprioritize documents retrieved using the IR relational system.

2. Prior Work

2.1. Implementation of an Information Retrieval System Using the Relational Model

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 inverted index stores each unique term or phrase from the collection and a list of all the documents containing each entry. The inverted index can also include frequency, offset, or other desired information. In the relational approach, this index is flattened or normalized and stored in a table. Queries can be implemented using standard structure query language (SQL) 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 our manual runs, we implemented an IR system Windows NT version of NCR/Teradata and Sybase/Adaptive Server Enterprise on Pentium SMP servers. This year's ad hoc and small web track submissions were run using an Oracle database system loaded on a SUN Solaris machine.

2.2. Relevance Feedback in the Relational Model

Our TREC-6 submission implemented automated relevance feedback within the relational model, using unchanged SQL. Prior work in relevance feedback has shown that this technique helps some queries, hurts some queries, but generally helps overall. The best results are achieved when a small set of terms is selected from the top documents [Lundquist97, Buckley95]. For instance, the terms might be ranked based on their normalized *idf* value and the least frequent terms added.

3. Implementation Details

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. We experimented with retrieval strategies, parser differences, and stemming/conflation for baseline improvements and then high-precision relevance feedback and thesaurus techniques for query expansion. What is most interesting about our work this year is the analysis and comparison of the best systems from TREC-7 and the new techniques we introduced in query expansion.

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 (including Singhal's pivoted normalized cosine measure.) We analyzed the result sets and found that they were very high in overlapping documents (relevant and nonrelevant) and in the ranking of those documents. We concluded that the different retrieval strategies (when based on tf*idf) do not account for the differences.

Token selection appears to have affected our effectiveness. We did not implement the GSL file that is typically used by OKAPI to conflate 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, our analysis indicated that stemming, phrase usage, and stop list differences were the main causes for variations in retrieval effectiveness. We experimented in all of these areas, but the key items to focus on are the 'stemming' and the title-phrase generation. We used the kstem+Porter equivalence groups to add term variants to the query [Allan98]. This 'stemming' was quite effective and resulted in 0.196 average precision recall for title+description. The new phrase generation technique creates new phrases of every pair-wise combination of title terms. The new phrases were minimally helpful on TREC-7 queries – getting us up to 0.20. The title-phrase technique did not cause serious degradation on any query and helped (although not by much) many queries. 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 so we moved on to query expansion.

3.2. Automatic Runs

3.2.1. 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, 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.

We analyzed the number of relevant documents returned in the top ten documents for each query. Interestingly, the top documents do not necessarily have to be relevant to be helpful. We found that when ten words were chosen from the top document returned, half of the time, the change in precision recall was opposite of what was expected. That is, when the top document was relevant, the average precision went down and when it was nonrelevant, the average precision went up. One explanation for this surprising result is that one document does not permit ranking words based on the number of top documents that contain the word. We examined the effect of relevance and nonrelevance of the top documents when using two documents for retrieval. We found that 34 queries brought back relevant documents in the top two. Of those, 22 improved and 12 degraded. This is much better than the 50% split when using only one document for feedback. However, one query actually improved even though its feedback was from nonrelevant documents.

To select terms, we used a modified Rocchio approach with the additional filter of requiring the term to occur in at least 2 of the top documents (N > 1.) These efforts increased our average precision recall to 0.2359. We note that similar work has been done earlier (most notably Mitra, Singhal and Buckley, SIGIR 1998 [Mitra98]) but our specific variations (automatic title concepts expanded with k-stems and N > 1) are new and effective.

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 ONE of the concepts. In order to limit the number of terms added in relevance feedback, we developed a filter which limits the terms added in relevance feedback to the top *x* terms, ranked by N**idf* where N > 1. N is the number of times the term occurs in the top 10 documents. As seen in Table 3.2.1-1, the best results were obtained by adding 10 terms. For this run, 23 queries improved and 14 degraded. Finally, we re-ranked our resulting set of documents by the percentage of query terms that were found in the document. This reranking gained a small improvement, bringing our final TREC-7 run to .2454.

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 3.2.1-1 Calibration of High-Precision Relevance Feedback Using top 10 documents

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

Table 3.2.1-1 Calibration of High-Precision Relevance Feedback Adding 10 terms

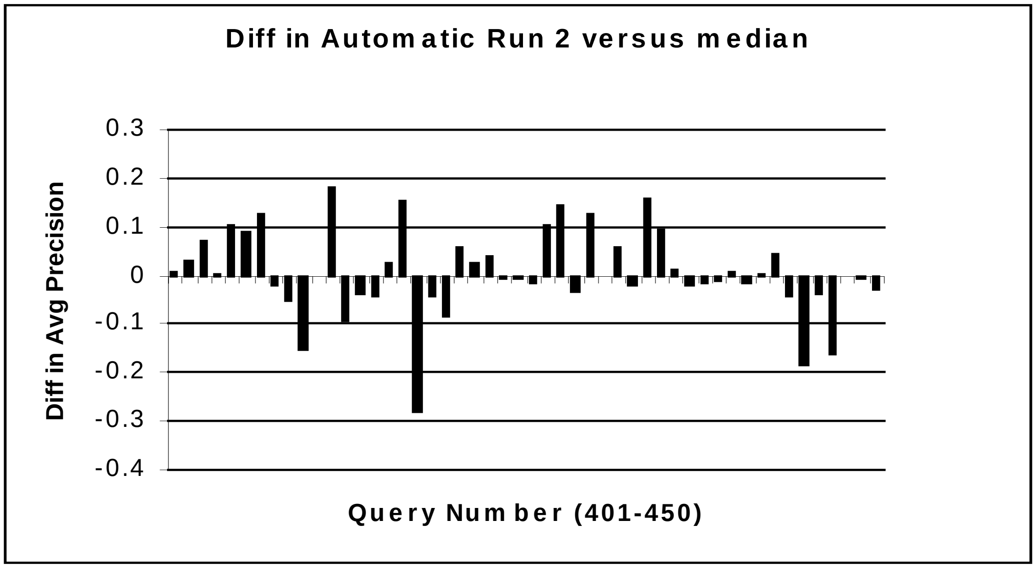

Figure 3.2.1-1: IIT Automatic Run-1 Difference from the Median

In summary, our automatic runs used title plus description portions of the topics and layered several techniques including a 570 word stop list, elimination of very frequent query terms such as ‘description’ and ‘relevant’, query expansion using porter+kstem equivalence classes, thresholding (eliminating) description terms with a document frequency greater than 2000, and pivoted normalized vector space similarity measure. Finally each run was reranked after retrieval using a function of the percentage of query terms present as the reranking criteria. Our second run (iit99au2) added a new high-precision relevance feedback pass. This run achieved an overall average precision recall of .2359 increasing to .2454 with the reranking.

3.3. Manual Run

3.3.1. Manual Run Implementation Details

Our approach to the manual ad hoc task for TREC-8 can best be described as the power of negative thinking. Consistent with previous years, our team used search, scoring and negation concepts. This year, 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. This technique eliminated many irrelevant documents from our results. For example, on Topic 447 — “Stirling engine” — we eliminated documents about *Stirling University* and people with the surname of *Stirling* which resulted in an average precision of 1.0.

Consistent with TREC-7, the IIT manual ad hoc queries used a set of search tokens consisting primarily of phrases and proper nouns. 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 Failure Analysis

IIT conducted failure analysis to determine why our manual ad hoc average precision was only at 0.41 even though we had manually scanned our answer sets prior to submission. We thought we had some pretty good result sets (of course we think this every year, but this year we spent more time reading the documents). 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 (which are summarized in Table 3.3.2-1).

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 3.3.2-1. Comparison of Relevance Judgments

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, even if the document did not specifically mention her name. Below are two examples that were judged non-relevant:

> *<num> Number: 423*
> *<title> Milosevic, Mirjana Markovic*
> *<desc> Description: Find references to Milosevic's wife, Mirjana Markovic.*
> *<narr> Narrative: Any mention of the Serbian president's wife is relevant, even if she is not named. She may be referred to by her nickname, Mira. A general mention of his family, without specifying his wife, is not relevant.*
>
> **Example #1**
> <DOCNO>FT942-13554</DOCNO> *taken from 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."
>
> **Example #2**
> < DOC NO> FBIS3-2 </DOCNO> *taken from text:*
> "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."

Topic 420 requests information on cases of carbon monoxide poisoning. However, the evaluation of relevance appears to be inconsistent. For example, two very similar documents are shown below – one judged relevant and one not.

> *<num> Number: 420 <title> carbon monoxide poisoning*
> *<desc> Description: How widespread is carbon monoxide poisoning on a global scale?*
> *<narr> Narrative: Relevant documents will contain data on what carbon monoxide poisoning is, symptoms, causes, and/or prevention. Advertisements for carbon monoxide protection products or services are not relevant. Discussions of auto emissions and air pollution are not relevant even though they can contain carbon monoxide.*

Here is an example of a document that NIST judged relevant, and we agree:

<DOCNO> LA010390-0086 </DOCNO> <TEXT>
Four people have died of **carbon monoxide poisoning** in a motor home outside a mountain lodge east of here, authorities said Tuesday. Sheriff's investigators said the victims died New Year's Day while camping in Mt. Laguna in the Cleveland National Forest. The coroner's office identified them as Katherine Walsh, 30; Michael McCrae,32; Conan Lemmer, 28, and Graham Rayner, 28, all of San Diego. The four were on a vacation in McCrae's motor home and were using a generator to power a heater in the 30-degree weather, but the exhaust pipe was too short to allow proper ventilation, deputies said. </TEXT>

However, the official judgment finds the following document irrelevant, and we disagree:

<DOCNO> LA121990-0004 </DOCNO> *taken from the text:*
". . .Tijuana officials confirmed Tuesday that three people are in custody in connection with the **carbon monoxide poisoning deaths** last week of 12 people in a house during a religious ceremony. ...As it turned out, Gloria Miranda Juarez and 11 others in the house died of carbon monoxide poisoning from a faulty lamp powered by butane gas, chemical experts with the judicial police said Monday. . ."

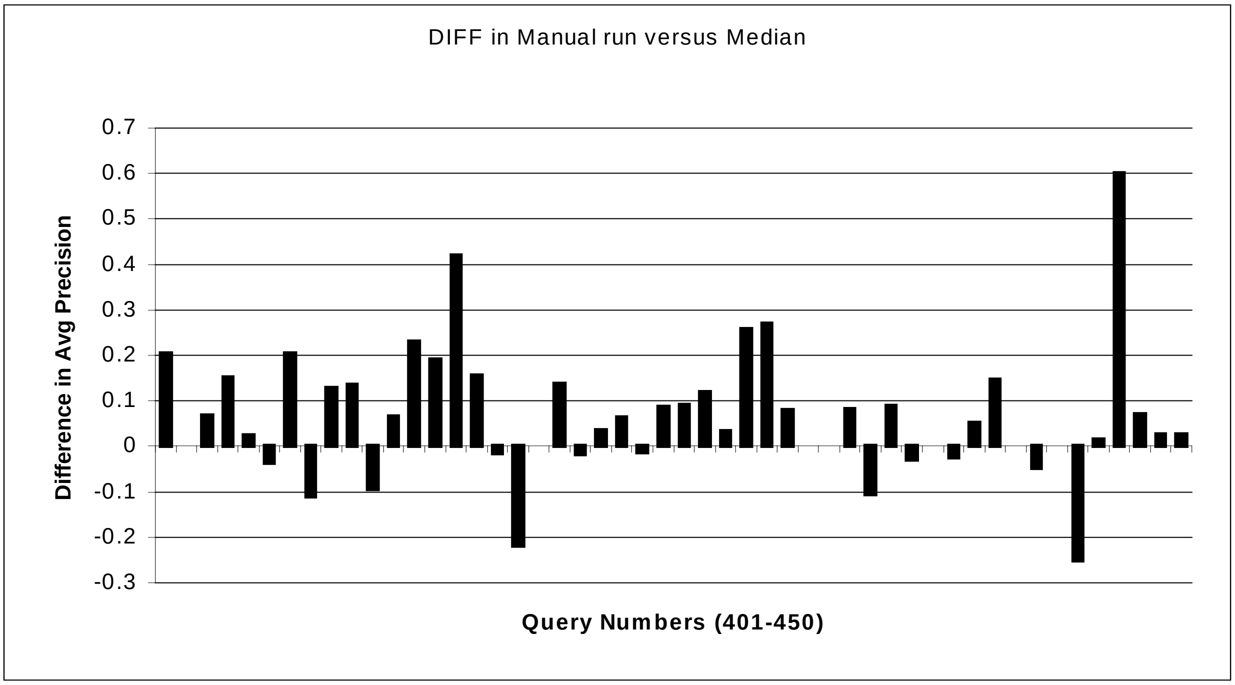

Figure 3.3.2-1: IIT Manual Run Difference from the median

The average precision for our manual run was officially scored at 0.4104. Correcting some of what may have only been clerical errors, would have enhanced our average precision.

3.4. Small Web Track

3.4.1. Small Web Track Implementation Details

Research for the Small Web Track was conducted on a Sun E10000 computer configured with 16 333mHz processors, 2 gigabytes of RAM, Solaris operating system, and Oracle 8 database management software.

Our Content-Only run (iit99wt1) did not attempt to reorder results based on web link information. The Link-Plus-Content runs (iit99wt2 and iit99wt3) began with the document sets retrieved during the Content-Only runs. While we made numerous initial efforts to incorporate link data and reorder documents based on links to or from other web pages, these 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.

The IR database system we used can be configured to retrieve documents beyond the 1000 documents submitted per query. We sought to use web links to identify and add documents to the solution set that were previously retrieved but not originally included in the Top 1000 solution set. The concept 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 solution set within the maximum 1000 documents per topic, the lowest ranking documents from the original Content-Only run were removed from the solution set. New documents were weighted and added to the retrieved documents solution set in such a manner that their original rankings were retained within the newly created solution set.

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

Table 3.4.1-1 IIT Small Web TREC-7 Benchmarks

3.4.2. Small Web Track Results

A comparison of results from our three small web track runs is found at Table 3.4.2-1.

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 3.4.2-1. IIT Small Web TREC-8 Results

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.

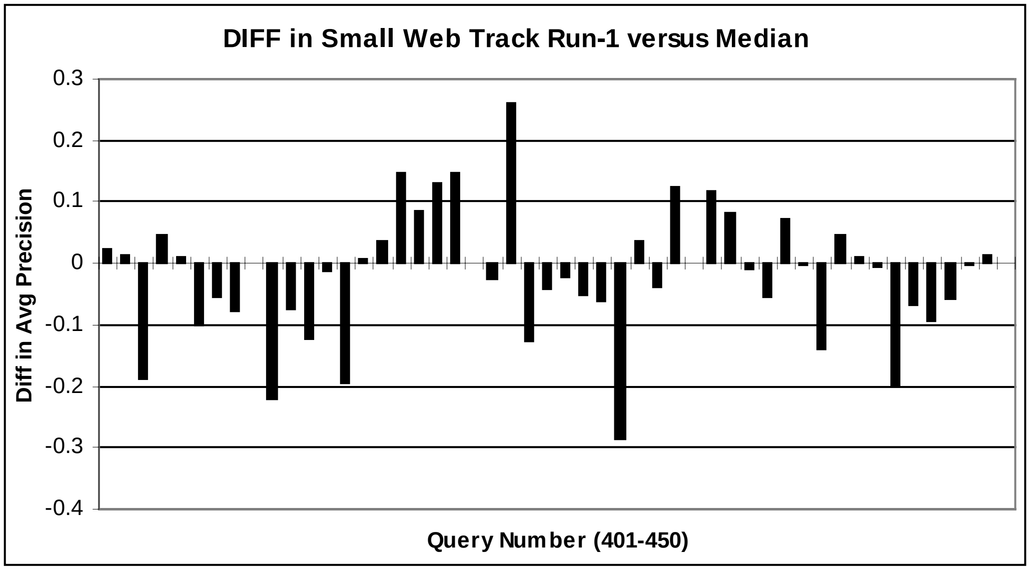

Figure 3.4.2-1: IIT Small Web Track Run-1 (Content-Only) Difference from the Median

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. Our average precision remained the same as the Content-Only run (at 0.2265).

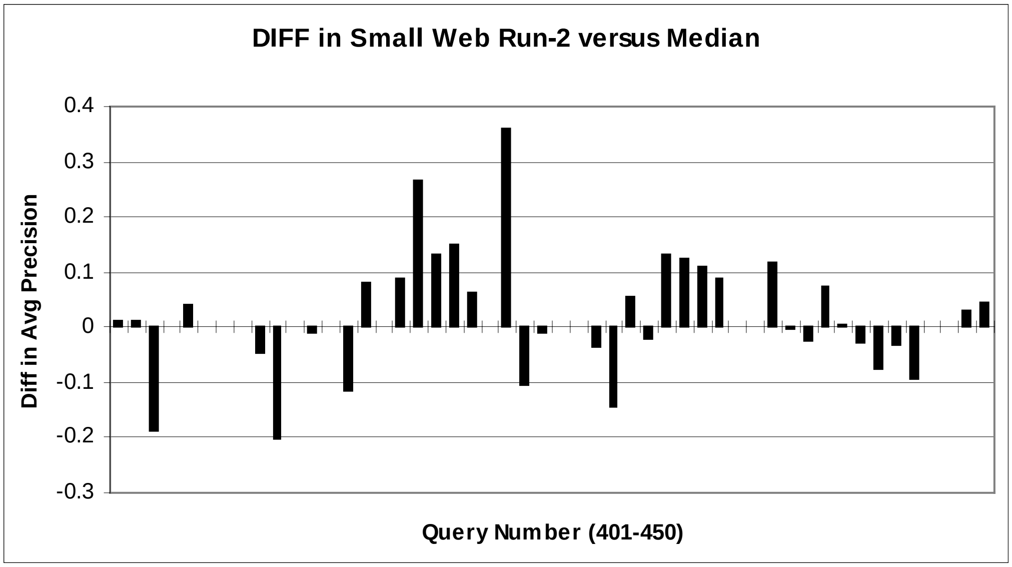

Figure 3.4.2-2: IIT Small Web Track Run-2 (Link-Plus-Content) Difference from the Median

Our third run (iit99wt3, a Link-Plus-Content run) was not evaluated and, therefore, could not contribute additional relevant documents to the qrels list. The average precision for Run-3 was slightly below the other two runs (with a value of 0.2264).

3.4.3. Small Web Track Failure Analysis

There are several factors that account for our performance in the small web track. First and foremost, is the fact that 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. Second, similar to the misgivings we experienced regarding several relevance judgments from the NIST document collection, we disagree with some of the relevance judgments assigned to the WT2g document collection. Rankings for our small web track submissions were based on title+description query analysis.

An additional contributing performance factor may be found in the TREC-8 small web track qrels set. The TREC-8 qrels set is only 35 percent as large as the TREC-7 qrels set (2279 vs. 6495). There are twice as many qrels for the ad hoc track, for example (4728 vs. 2279). Only 5 (10%) of the small web track topics have over 100 relevant documents; for the ad hoc track 20 (40%) of the topics have more than 100 relevant documents. Over two-thirds (34 out of 50) of the WT2g topics have 50 or fewer relevant documents, as opposed to 19 ad hoc qrels that have 50 or fewer qrels. This information is summarized in Table 3.4.3-1.

	QRELS per Topic					
TREC-8 Track	**1-25**	**26-50**	**51-75**	**76-100**	**101-200**	**>200**
Ad Hoc	11	8	8	3	**15**	**5**
Small Web	**17**	**17**	7	4	5	0

Table 3.4.3-1. TREC-8 Small Web Track vs. Ad Hoc Number of QRELS per Topic

4. Results

Table 4-1 summarizes the results of our 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 4-1: IIT TREC-8 Results Summary

5. Conclusions and Future Work

For TREC-8, we focused on improving our baseline system and then introducing some new feedback techniques. The bottom line is that we conducted numerous experiments and analysis this year, and were able to identify some key enhancements to our parser as well are our feedback engine. We introduced a technique for using k-stem conflation groups (based on our prior success with manual tracks in prior years) to expand title-term concepts and use this as a filter for high-precision relevance feedback.

Clearly, at least for manual ad hoc, phrases and nouns are important elements in runs with high average precision. One of our future challenges remains the automation of the techniques used to add high quality phrases into our search engine.

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.

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.

6. 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 Twentyfirst 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.