Experiments with the Negotiated Boolean Queries of the TREC 2007 Legal Discovery Track

Stephen Tomlinson
Open Text Corporation
Ottawa, Ontario, Canada
stomlins@opentext.com
http://www.opentext.com/

February 5, 2008

Abstract

We analyze the results of several experimental runs submitted for the TREC 2007 Legal Track (also sometimes known as the Legal Discovery Track). We submitted 4 boolean query runs (the initial proposal by the defendant, the rejoinder by the plaintiff, the final negotiated query, and a variation of the final query which had proximity distances doubled). We submitted 2 vector query runs (one based on the keywords of the final negotiated query, and another based on the (natural language) request text). We submitted a blind feedback run based on the final negotiated boolean query. Finally, we submitted a fusion run of the final boolean, request text and final vector runs. We found that none of the runs had a higher mean estimated Recall@B than the original final negotiated boolean query.

1 Introduction

Livelink ECM - eDOCS SearchServer^{TM} is a toolkit for developing enterprise search and retrieval applications. The SearchServer kernel is also embedded in other components of the Livelink ECM - eDOCS Suite[1].

SearchServer works in Unicode internally [5] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (TREC [9], CLEF [4] and NTCIR [7]) have provided judged test collections for objective experimentation with SearchServer in more than a dozen languages.

This paper describes experimental work with SearchServer (experimental post-6.0 builds) for legal discovery.

2 Legal Discovery Track

In the Legal Discovery Track (also known as the TREC Legal Track), the collection to be searched was the IIT Complex Document Information Processing (IIT CDIP) test collection [6]. It contained 6,910,192 metadata records from US tobacco companies; 6,794,895 of the records included document text of varying quality from an optical character reader. Uncompressed, the collection was 61,251,357,065 bytes (57.0 GB). The average record size (including metadata markup and the ocr document) was 8864 bytes.

[1] Livelink, Open Text^{TM} and SearchServer^{TM} are trademarks or registered trademarks of Open Text Corporation in the United States of America, Canada, the European Union and/or other countries. This list of trademarks is not exhaustive. Other trademarks, registered trademarks, product names, company names, brands and service names mentioned herein are property of Open Text Corporation or other respective owners.

In legal discovery, the goal is to return all documents responsive (relevant) to a production request. For the main task (also known as the ad hoc task) of the TREC Legal Track, the organizers created 50 new production requests (topics), numbered from 52 to 101. Each topic included a "request text" (a natural language description of the request, typically one-sentence), a "defendant query" (an initial boolean query proposed by the defendant), a "plaintiff query" (a rejoinder boolean query from the plaintiff) and a "final negotiated query" (the final boolean query from the negotiations). (Examples appear below.) Note that last year the plaintiff and final queries were the same, but this year they could differ (hence this year there were 3 boolean queries per topic instead of just 2). Also, this year the final boolean query was known to always match between 100 and 25,000 records, whereas last year the range was larger and not announced in advance. During the assessing phase, 7 topics were dropped, leaving 43 topics.

[13] has more details on the track and task, and [1] and [2] have more background on legal discovery in general. Also, background on last year's track is in [12] and [3].

2.1 Indexing

Our index included both the metadata and the ocr document of each record. We indexed from the "</tid>" tag to the "</record>" tag, which meant both the metadata and the ocr document were in the FT_TEXT column. Any tags themselves were indexed (we just didn't bother to discard them). Entities (e.g. "&amp;") were converted to the character they represented (e.g. "&").

We did not use a stopword list, and we also indexed most punctuation as 1-character words (exceptions were the hyphen and apostrophe, which were treated as 1-character stopwords). The contents of the "<dd>" section of the metadata were additionally indexed in a separate DOCDATE column, though this column was not used by the queries this year.

The index supported both searching on just the surface forms of the words and also searching on inflections from English lexical stemming. The documents were assumed to be in the Windows-1252 character set when converted to Unicode. Words were normalized to upper-case and any accents were dropped.

2.2 Searching

The techniques used for the 8 submitted runs of July 2007 are described below. The relevance ranking approach was the same for all runs. The relevance function dampened the term frequency and adjusted for document length in a manner similar to Okapi [8] and dampened the inverse document frequency using an approximation of the logarithm. For wildcard terms (e.g. "televis!"), all variants (e.g. "television", "televised", "televisions", etc.) were treated as occurrences of the same term for term frequency purposes, and inverse document frequency was based on the most common variant. For runs which used inflectional matching, these calculations were based on the stems of the terms. For terms in phrases or proximity contraints of boolean queries, only occurrences of the term satisfying the phrase or proximity counted towards term frequency.

The 8 submitted runs were as follows:

otL07fb (final boolean run): The submitted otL07fb run used the final negotiated query, respecting the boolean operators such as AND, phrase, proximity, NOT, etc. Full wildcard matching was supported. Relevance-ranking was still used to order the matching rows. The run was labelled as manual because some hand-editing was done to convert the queries to the SearchSQL syntax of SearchServer, but the run was automatic in spirit because it just implemented the final boolean query intended by the negotiators. For example, for topic 74, for which the final negotiated query was "`(scien! OR stud! OR research) AND ("air quality" w/15 health)`", the corresponding SearchSQL statement was

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM LEGAL07FULL
WHERE ((FT_TEXT CONTAINS 'scien%', 'stud%', 'research')
 AND   (FT_TEXT CONTAINS 'air quality' within 15 words of 'health'))
ORDER BY REL DESC;
```

otL07db (defendant boolean run): The submitted otL07db run was the same as otL07fb except that the defendant query was used instead of the final negotiated boolean query. For example, for topic 74, for which the initial query proposed by the defendant was “"health effect!" w/10 "air quality"”, the WHERE clause of the corresponding SearchSQL statement was

```
WHERE (FT_TEXT CONTAINS 'health effect%' within 10 words of 'air quality')
```

otL07pb (plaintiff boolean run): The submitted otL07pb run was the same as otL07fb except that the plaintiff query was used instead of the final negotiated boolean query. For example, for topic 74, for which the rejoinder query from the plaintiff was “(scien! OR stud! OR research) AND ("air quality" OR health)”, the WHERE clause of the corresponding SearchSQL statement was

```
WHERE ((FT_TEXT CONTAINS 'scien%', 'stud%', 'research')
 AND   (FT_TEXT CONTAINS 'air quality', 'health'))
```

Note: for some topics, the plaintiff query matched more than 25,000 records, in which case the lower-ranking matches were omitted because of the depth-25,000 submission limit.

otL07fb2x (twice-distance boolean run): The submitted otL07fb2x run was the same as otL07fb except that any word proximity distances in the query were doubled (phrases, however, were not adjusted). For example, for topic 74, the “w/15” was interpreted as “w/30”, hence the WHERE clause of the corresponding SearchSQL statement was

```
WHERE ((FT_TEXT CONTAINS 'scien%', 'stud%', 'research')
 AND   (FT_TEXT CONTAINS 'air quality' within 30 words of 'health'))
```

otL07fv (final vector run): The submitted otL07fv run was the same as otL07fb except that the boolean operators such as AND, phrases and proximities were dropped (all operators became an OR), and punctuation was dropped. Full wildcarding was still respected. For example, for topic 74, the WHERE clause of the corresponding SearchSQL statement was

```
WHERE (FT_TEXT CONTAINS 'scien%'|'stud%'|'research'|'air'|'quality'|'health')
```

otL07rvl (request text vector run): The submitted otL07rvl run was the same as otL07fv except that (1) the terms were taken from the request text field instead of the final negotiated query field, (2) linguistic expansion from English inflectional stemming was applied, and (3) common instruction words (e.g. “please”, “produce”, “documents”) were manually removed. For example, for topic 74, for which the request text was “All scientific studies expressly referencing health effects tied to indoor air quality.”, the WHERE clause of the corresponding SearchSQL statement was

```
WHERE FT_TEXT CONTAINS 'scientific'|'studies'|'expressly'|
         'health'|'effects'|'tied'|'indoor'|'air'|'quality'
```

otL07frw (fusion run): The submitted otL07frw run was a weighted fusion of the final boolean, request text vector and final vector runs: weight 3 on otL07fb, weight 2 on otL07rvl, weight 1 on otL07fv. (Each input run was retrieved to depth 25000 (or B in the case of the final boolean run). The relevance score (also known as the retrieval status value or rsv score) for each document was multiplied by the applicable weight, and the resulting scores for each document were added together and used as the basis for the fusion ranking.)

otL07fbe (blind feedback run): The submitted otL07fbe run was a blind feedback run based 50% on otL07fb and 25% each on expansion queries from the first 2 rows of otL07fb.

For each run, only 25000 rows were allowed to be submitted for each query.

2.3 Results

Table 1 lists several mean scores for the 8 submitted runs. The retrieval measures are defined in Section 3.1 of the Glossary at the end of the paper. The highest mean scores of each measure are in bold; however, see Tables 2-4 for which mean differences are statistically significant. (The columns of Tables 2-4 are explained in Section 3.2 of the Glossary.)

The next 7 sections look at the differences between the submitted runs in more detail.

Table 1: Mean Scores of Submitted Ad Hoc Task Runs

Run	Est. R@B	Est. R25000	Est. P5	Est. P@B	Est. P25000	Est. Gray@B	S1J	GS10J	(raw) R-Prec
otL07fb	**0.216**	0.216	0.507	**0.292**	0.056	**0.042**	24/43	0.863	0.201
otL07fb2x	0.209	0.242	0.486	0.282	0.065	0.031	21/43	0.837	0.193
otL07frw	0.193	**0.428**	**0.550**	0.278	0.150	0.015	**26/43**	**0.883**	**0.224**
otL07pb	0.186	0.327	0.424	0.235	0.119	0.025	17/43	0.792	0.147
otL07fbe	0.169	0.369	0.492	0.273	**0.160**	0.015	22/43	0.792	0.178
otL07fv	0.163	0.357	0.467	0.225	0.129	0.005	20/43	0.856	0.176
otL07rvl	0.153	0.420	0.471	0.235	0.151	0.010	15/43	0.850	0.187
otL07db	0.027	0.027	0.301	0.026	0.006	0.003	15/43	0.576	0.074

2.3.1 Defendant Boolean vs. Final Boolean

The initial proposal by the defendant always matched fewer than B documents, and it always matched fewer relevant documents than the final boolean query (except for topic 77, which was a tie because neither query found a relevant document).

The defendant boolean also had relatively low Precision@*k* scores because it often did not retrieve to the full depth (*k*=5, B or 25000) at which precision was measured. In fact, for 11 of the original 50 topics, the defendant boolean had zero matches.

If you measure precision over the set of retrieved documents instead of to a fixed depth *k*, i.e. use the "est_P_set" measure in the l07_eval output instead of "est_P5", "est_PB" or "est_P25000", then the defendant boolean query had a higher precision than the final boolean query for 20 of the 43 queries. Again, for some of the other 23 queries the defendant boolean query did not retrieve any documents at all (hence set precision is undefined, though the l07_eval utility assigned a zero score for these cases).

To summarize, the final boolean query always had the same or higher recall as the defendant boolean query, but the defendant boolean query often (though not always) had the higher precision.

2.3.2 Plaintiff Boolean vs. Final Boolean

The rejoinder query from the plaintiff usually matched more than B documents (the 3 exceptions were topics 66, 68 and 99, and topic 68 was one of the discarded topics). Hence it is not surprising that the plaintiff query had a higher Recall@25000 than the final boolean query (as per the "pb-fb" entry for R25000 in Table 3). But the final boolean query had a higher mean Recall@B (though this difference was not statistically significant as per the "pb-fb" entry for R@B in Table 2).

Note that many of the plaintiff queries matched more than 25,000 documents, so not all plaintiff matches could be submitted. The highest number of matches was 858,700 for topic 74.

Table 2 shows that the plaintiff queries led to a statistically significant decline in mean Precision@B compared to the final boolean queries (as per the "pb-fb" entry for P@B in Table 2), even though the table also shows that the plaintiff query had the higher P@B for 16 of the 43 topics.

Apparently the plaintiff sacrificed a lot of relevant documents in the negotiations, on average, as the final boolean query averaged 22% recall, while the plaintiff boolean query averaged more than 32% recall (perhaps a lot more since some plaintiff queries had more than 25,000 matches). We suspect the plaintiff didn't realize this many relevant documents were being given up. Furthermore, the plaintiff presumably would be unhappy that the final boolean query only had 22% recall on average.

2.3.3 Impact of Doubling Proximity Distances

Doubling the proximity distances in the final boolean query would of course potentially match more documents. R@25000 was increased for 22 of the topics, with no declines (as per the "fb2x-fb" entry of Table

Table 2: Impact of Legal Discovery Techniques (R@B, P@B and Gray@B)

Expt	ΔR@B	95% Conf	vs.	3 Extreme Diffs (Topic)
fb2x-fb	−0.007	(−0.027, 0.013)	14-8-21	−0.39 (72), −0.08 (58), 0.05 (61)
frw-fb	−0.023	(−0.074, 0.028)	20-20-3	−0.84 (58), −0.40 (72), 0.21 (65)
pb-fb	−0.029	(−0.090, 0.031)	15-27-1	−0.76 (58), −0.48 (72), 0.46 (63)
fbe-fb	−0.047	(−0.115, 0.021)	16-21-6	−0.93 (58), −0.82 (84), 0.54 (90)
fv-fb	−0.053	(−0.122, 0.017)	16-26-1	−0.86 (58), −0.75 (72), 0.52 (63)
rvl-fb	−0.063	(−0.126, 0.000)	15-28-0	−0.86 (58), −0.75 (72), 0.30 (65)
db-fb	−0.189	(−0.266,−0.112)	0-42-1	−0.98 (52), −0.92 (58), 0.00 (77)
db-pb	−0.159	(−0.224,−0.094)	1-40-2	−0.96 (52), −0.68 (63), 0.00 (80)
frw-rvl	0.040	(0.013, 0.068)	33-10-0	0.34 (72), 0.32 (61), −0.09 (78)
frw-fv	0.030	(−0.013, 0.073)	30-11-2	0.46 (61), 0.40 (65), −0.41 (63)
rvl-fv	−0.010	(−0.053, 0.032)	19-20-4	−0.62 (63), −0.23 (85), 0.49 (65)
fbe-rvl	0.016	(−0.032, 0.065)	19-23-1	0.55 (72), 0.51 (90), −0.36 (84)
fbe-fv	0.006	(−0.057, 0.068)	21-21-1	−0.73 (63), 0.48 (65), 0.55 (72)
fbe-frw	−0.024	(−0.066, 0.017)	16-24-3	0.49 (90), −0.32 (63), −0.46 (84)
	ΔP@B			
fb2x-fb	−0.011	(−0.028, 0.007)	10-15-18	0.17 (94), −0.15 (72), −0.16 (74)
frw-fb	−0.014	(−0.040, 0.012)	19-21-3	−0.35 (53), −0.15 (72), 0.11 (70)
pb-fb	−0.057	(−0.099,−0.015)	16-26-1	−0.49 (80), −0.39 (87), 0.16 (59)
fbe-fb	−0.019	(−0.062, 0.025)	14-26-3	−0.41 (73), 0.29 (71), 0.35 (79)
fv-fb	−0.067	(−0.118,−0.017)	17-25-1	−0.45 (79), −0.40 (72), 0.35 (98)
rvl-fb	−0.057	(−0.098,−0.016)	19-24-0	−0.53 (53), −0.40 (72), 0.14 (80)
db-fb	−0.266	(−0.336,−0.196)	1-41-1	−0.97 (69), −0.70 (71), 0.01 (65)
db-pb	−0.209	(−0.279,−0.139)	1-40-2	−0.98 (69), −0.67 (57), 0.04 (65)
frw-rvl	0.043	(0.014, 0.072)	31-12-0	0.25 (72), 0.25 (87), −0.18 (80)
frw-fv	0.054	(0.005, 0.102)	26-16-1	0.47 (73), 0.41 (79), −0.29 (98)
rvl-fv	0.011	(−0.035, 0.056)	21-21-1	0.41 (80), 0.38 (73), −0.36 (53)
fbe-rvl	0.038	(−0.027, 0.103)	18-25-0	0.62 (53), 0.59 (72), −0.43 (73)
fbe-fv	0.049	(−0.020, 0.117)	22-20-1	0.81 (79), 0.59 (72), −0.52 (98)
fbe-frw	−0.005	(−0.056, 0.046)	12-27-4	−0.52 (73), 0.40 (79), 0.44 (53)
	ΔG@B			
fb2x-fb	−0.011	(−0.022,−0.001)	1-13-29	−0.17 (63), −0.08 (72), 0.00 (89)
frw-fb	−0.027	(−0.046,−0.009)	3-23-17	−0.25 (86), −0.20 (87), 0.01 (56)
pb-fb	−0.017	(−0.044, 0.009)	9-22-12	−0.25 (86), −0.20 (87), 0.18 (61)
fbe-fb	−0.027	(−0.045,−0.009)	2-24-17	−0.25 (86), −0.20 (87), 0.00 (71)
fv-fb	−0.037	(−0.059,−0.015)	5-26-12	−0.25 (86), −0.20 (87), 0.08 (79)
rvl-fb	−0.033	(−0.057,−0.008)	4-26-13	−0.25 (86), −0.20 (87), 0.20 (89)
db-fb	−0.040	(−0.062,−0.017)	3-26-14	−0.25 (86), −0.23 (69), 0.00 (101)
db-pb	−0.022	(−0.040,−0.004)	3-16-24	−0.22 (69), −0.18 (61), 0.08 (72)
frw-rvl	0.005	(−0.010, 0.021)	15-4-24	−0.20 (89), 0.14 (56), 0.15 (77)
frw-fv	0.010	(−0.003, 0.022)	14-8-21	0.15 (77), 0.14 (56), −0.08 (79)
rvl-fv	0.004	(−0.004, 0.013)	9-9-25	0.16 (89), 0.04 (69), −0.01 (72)
fbe-rvl	0.005	(−0.010, 0.021)	12-10-21	−0.20 (89), 0.14 (69), 0.15 (77)
fbe-fv	0.010	(−0.004, 0.023)	11-9-23	0.19 (69), 0.15 (77), −0.08 (79)
fbe-frw	0.000	(−0.006, 0.007)	5-11-27	0.11 (69), −0.02 (98), −0.06 (56)

Table 3: Impact of Legal Discovery Techniques (R25000 and P25000)

Expt	ΔR25000	95% Conf	vs.	3 Extreme Diffs (Topic)
fb2x-fb	0.026	(0.003, 0.049)	22-0-21	0.36 (59), 0.33 (61), 0.00 (101)
frw-fb	0.212	(0.109, 0.316)	35-8-0	0.96 (87), 0.93 (60), −0.83 (58)
pb-fb	0.112	(0.036, 0.187)	35-6-2	0.75 (59), 0.70 (60), −0.70 (58)
fbe-fb	0.153	(0.051, 0.255)	33-8-2	0.94 (87), 0.94 (79), −0.87 (58)
fv-fb	0.141	(0.051, 0.231)	32-11-0	0.87 (75), 0.73 (63), −0.78 (58)
rvl-fb	0.204	(0.099, 0.308)	33-10-0	0.96 (87), 0.93 (60), −0.84 (58)
db-fb	−0.189	(−0.266,−0.112)	0-42-1	−0.98 (52), −0.92 (58), 0.00 (77)
db-pb	−0.300	(−0.391,−0.209)	0-42-1	−1.00 (52), −0.85 (61), 0.00 (77)
frw-rvl	0.009	(−0.008, 0.026)	21-5-17	0.19 (100), 0.16 (78), −0.18 (92)
frw-fv	0.072	(0.013, 0.130)	22-19-2	0.87 (87), 0.47 (77), −0.22 (84)
rvl-fv	0.063	(0.002, 0.123)	21-22-0	0.86 (87), 0.47 (77), −0.27 (84)
fbe-rvl	−0.051	(−0.148, 0.047)	16-27-0	−0.92 (60), −0.87 (75), 0.70 (89)
fbe-fv	0.012	(−0.096, 0.121)	17-23-3	0.93 (79), 0.84 (87), −0.87 (75)
fbe-frw	−0.059	(−0.157, 0.038)	12-29-2	−0.92 (60), −0.87 (75), 0.68 (89)
	ΔP25000			
fb2x-fb	0.009	(0.001, 0.017)	24-5-14	0.13 (94), 0.09 (69), −0.01 (65)
frw-fb	0.094	(0.041, 0.147)	36-7-0	0.80 (69), 0.62 (80), −0.05 (97)
pb-fb	0.062	(0.022, 0.103)	35-7-1	0.58 (69), 0.46 (98), −0.06 (99)
fbe-fb	0.103	(0.038, 0.168)	34-8-1	0.80 (69), 0.76 (71), −0.05 (94)
fv-fb	0.072	(0.026, 0.119)	33-10-0	0.80 (69), 0.30 (53), −0.06 (95)
rvl-fb	0.094	(0.040, 0.149)	34-9-0	0.80 (69), 0.62 (80), −0.09 (95)
db-fb	−0.050	(−0.079,−0.021)	1-41-1	−0.55 (74), −0.22 (95), 0.00 (65)
db-pb	−0.113	(−0.168,−0.057)	1-41-1	−0.78 (69), −0.55 (74), 0.01 (65)
frw-rvl	−0.001	(−0.011, 0.010)	16-12-15	−0.13 (92), −0.09 (53), 0.08 (74)
frw-fv	0.021	(−0.005, 0.048)	26-14-3	0.36 (80), 0.22 (77), −0.13 (67)
rvl-fv	0.022	(−0.004, 0.048)	25-17-1	0.36 (80), 0.19 (77), −0.11 (96)
fbe-rvl	0.009	(−0.041, 0.059)	20-23-0	0.54 (82), 0.40 (96), −0.46 (80)
fbe-fv	0.031	(−0.012, 0.074)	21-21-1	0.56 (71), 0.45 (82), −0.21 (53)
fbe-frw	0.009	(−0.039, 0.058)	20-23-0	0.54 (82), 0.40 (96), −0.46 (80)

3).

However, the resulting mean R@B and P@B were lower when doubling the proximity distances (though these declines were not statistically significant as per the "fb2x-fb" entries in Table 2). Some other precision-based measures were borderline significantly lower when doubling the proximity distances, e.g. the GS10J and (raw) R-Prec measures (as per the "fb2x-fb" entries of Table 4).

Note: if the judgements were complete, R@B and P@B would favor the same run for each topic (though the mean scores could favor different runs). But because of the way that R@B and P@B are estimated from sampling, it's not always the case that the same run is favored by both the R@B and P@B estimates for the same topic. e.g. Table 2 shows that for R@B, the fb2x run scored higher on 14, lower on 8 and tied 21, whereas for P@B, the fb2x run scored higher on 10, lower on 15 and tied 18. Hence it's good to check both the estimated R and P measures at each depth.

Besides our interest in investigating if the negotiators picked reasonable proximity distances, another motivating reason for conducting this experiment was that the proximity distances in our implementation may be tighter than one would intuitively expect because most punctuation characters counted as words (as a side effect of enabling punctuation indexing). This run hence may have added documents to the pool that might be considered matches to the original final boolean query by some other implementations (particularly

Table 4: Impact of Legal Discovery Techniques (GS10J, P5 and R-Prec)

Expt	ΔGS10J	95% Conf	vs.	3 Extreme Diffs (Topic)
fb2x-fb	−0.025	(−0.050, 0.000)	2-9-32	−0.32 (64), −0.30 (89), 0.13 (85)
frw-fb	0.021	(−0.009, 0.051)	11-5-27	0.32 (101), 0.27 (75), −0.27 (64)
pb-fb	−0.070	(−0.149, 0.008)	8-16-19	−0.95 (60), −0.86 (99), 0.34 (75)
fbe-fb	−0.071	(−0.123,−0.018)	5-13-25	−0.73 (97), −0.61 (59), 0.07 (98)
fv-fb	−0.007	(−0.071, 0.058)	11-9-23	−0.77 (64), −0.57 (78), 0.43 (86)
rvl-fb	−0.013	(−0.092, 0.067)	12-18-13	−0.88 (64), 0.58 (86), 0.73 (73)
db-fb	−0.287	(−0.426,−0.147)	8-24-11	−1.00 (56), −1.00 (83), 0.32 (101)
db-pb	−0.216	(−0.369,−0.063)	13-21-9	−1.00 (56), −1.00 (70), 0.95 (60)
frw-rvl	0.034	(−0.036, 0.103)	16-8-19	−0.73 (73), −0.58 (86), 0.62 (64)
frw-fv	0.027	(−0.024, 0.078)	14-10-19	0.50 (64), 0.50 (78), −0.43 (86)
rvl-fv	−0.006	(−0.070, 0.057)	14-18-11	0.72 (73), 0.50 (78), −0.63 (85)
fbe-rvl	−0.058	(−0.162, 0.045)	15-16-12	−0.98 (59), −0.73 (73), 0.88 (64)
fbe-fv	−0.064	(−0.158, 0.029)	11-16-16	−0.98 (59), −0.77 (75), 0.77 (64)
fbe-frw	−0.092	(−0.162,−0.021)	5-18-20	−0.73 (97), −0.72 (59), 0.27 (64)
	ΔP5			
fb2x-fb	−0.021	(−0.056, 0.015)	3-6-34	−0.50 (89), −0.20 (94), 0.20 (85)
frw-fb	0.043	(−0.023, 0.109)	11-10-22	0.80 (101), 0.60 (67), −0.20 (66)
pb-fb	−0.083	(−0.181, 0.015)	10-14-19	−1.00 (65), −0.80 (87), 0.60 (67)
fbe-fb	−0.016	(−0.077, 0.046)	8-12-23	0.60 (76), −0.40 (58), −0.50 (56)
fv-fb	−0.040	(−0.135, 0.055)	10-18-15	−0.80 (65), −0.67 (78), 0.80 (96)
rvl-fb	−0.036	(−0.147, 0.074)	15-17-11	−0.80 (95), −0.80 (83), 0.75 (101)
db-fb	−0.207	(−0.344,−0.069)	8-25-10	−1.00 (71), −1.00 (69), 0.80 (97)
db-pb	−0.124	(−0.272, 0.025)	12-22-9	−1.00 (69), −1.00 (67), 1.00 (65)
frw-rvl	0.079	(−0.017, 0.176)	19-10-14	0.80 (95), 0.80 (58), −0.60 (98)
frw-fv	0.083	(−0.017, 0.183)	15-9-19	0.80 (65), 0.80 (58), −0.60 (80)
rvl-fv	0.003	(−0.104, 0.111)	16-15-12	−0.80 (95), −0.80 (85), 0.60 (98)
fbe-rvl	0.021	(−0.092, 0.134)	14-15-14	0.80 (95), 0.80 (83), −0.75 (101)
fbe-fv	0.024	(−0.068, 0.117)	15-11-17	0.80 (76), 0.60 (65), −0.60 (90)
fbe-frw	−0.059	(−0.143, 0.026)	12-14-17	−0.80 (58), −0.80 (101), 0.40 (100)
	ΔRPrec			
fb2x-fb	−0.008	(−0.017, 0.000)	6-13-24	−0.09 (63), −0.09 (72), 0.04 (74)
frw-fb	0.023	(0.007, 0.039)	24-6-13	0.17 (96), 0.12 (70), −0.10 (60)
pb-fb	−0.055	(−0.094,−0.015)	14-24-5	−0.51 (87), −0.33 (65), 0.15 (96)
fbe-fb	−0.023	(−0.053, 0.007)	18-17-8	−0.45 (63), −0.27 (97), 0.16 (96)
fv-fb	−0.025	(−0.064, 0.014)	20-19-4	−0.38 (65), −0.27 (63), 0.20 (98)
rvl-fb	−0.014	(−0.057, 0.028)	22-19-2	−0.36 (63), −0.30 (83), 0.26 (89)
db-fb	−0.127	(−0.168,−0.087)	3-36-4	−0.51 (87), −0.41 (83), 0.16 (97)
db-pb	−0.073	(−0.121,−0.024)	10-27-6	−0.44 (52), −0.41 (83), 0.32 (97)
frw-rvl	0.037	(−0.002, 0.076)	24-16-3	0.30 (83), 0.27 (63), −0.28 (89)
frw-fv	0.048	(0.009, 0.087)	24-14-5	0.45 (65), 0.32 (84), −0.22 (89)
rvl-fv	0.011	(−0.021, 0.043)	23-11-9	0.40 (65), −0.20 (52), −0.22 (95)
fbe-rvl	−0.008	(−0.052, 0.035)	17-25-1	−0.38 (97), 0.27 (83), 0.28 (66)
fbe-fv	0.002	(−0.041, 0.046)	15-24-4	0.48 (65), 0.28 (66), −0.30 (97)
fbe-frw	−0.045	(−0.075,−0.016)	7-27-9	−0.36 (63), −0.31 (97), 0.12 (90)

those which do not count punctuation characters as words).

2.3.4 Boolean vs. Vector

The decline in mean P@B from dropping the boolean operators was statistically significant this year (as per the "fv-fb" entry in Table 2). Mean R@B also fell, but the decline was not statistically significant. We look at the topics listed in the "3 Extreme Diffs" column of Table 2 for the R@B "fv-fb" entry:

Topics on which the boolean query scored higher in R@B:

- Topic 58 ("Please produce any and all documents that discuss health problems caused by HPF, including, but not limited to immune disorders, toxic myopathy, chronic fatigue syndrome, liver dysfunctions, irregular heart-beat, reactive depression, and memory loss"): This topic had the largest difference in R@B (and also R@25000) in favour of the boolean run (as per the "fv-fb" entries of Tables 2 and 3). For this query, B was 8183 and the estimated number of relevant documents was 1151. The recall of the boolean query was 94%, while the vector query's Recall@B was just 8%. The final boolean query (`Phosphat! w/75 (caus! OR relat! OR assoc! OR derive! OR correlat!) w/75 (health OR disorder! OR toxic! OR "chronic fatigue" OR dysfunction! OR irregular OR memor! OR immun! OR myopath! OR liver! OR kidney! OR heart! OR depress! OR loss OR lost)`) required that a term starting with "phosphat" be in the document (and furthermore be near a term or phrase indicative of a health problem). A lot of the top-ranked matches for the vector query, however, did not have a term starting with "phosphat", as it was just 1 of 21 terms in the vector form of the query.

- Topic 72 ("All documents referring to the scientific or chemical process(es) which result in onions have the effect of making persons cry"): This topic had the 2nd-largest difference in R@B (and also P@B) in favour of the boolean run (as per the "fv-fb" entries of Table 2). For this query, B was 119 and the estimated number of relevant documents was 98. The recall of the boolean query was 78%, while the vector query's Recall@B was just 3%. The final boolean query (`((scien! OR research! OR chemical) w/25 onion!) AND (cries OR cry! OR tear!)`) matched some long documents (e.g. brq10e00 which was 105 pages) with just one passing reference to the chemical properties of onions, which were judged relevant. A lot of the top-ranked matches for the vector query did not contain the term 'onion' (or 'onions'), but had lots of occurrences of other query terms.

A topic on which the vector query scored higher in R@B:

- Topic 63 ("Please produce any and all documents that specifically discuss an exclusivity clause in a sugar contract"): This topic had the largest difference in R@B (and 2nd largest in R@25000) in favour of the vector run (as per the "fv-fb" entries of Table 2). For this query, B was 294 and the estimated number of relevant documents was 18.6 (the fewest of any topic). The recall of the boolean query was 27%, while the vector query's Recall@B was 79%. The final boolean query (`(Sugar w/20 (contract! OR agreement! OR deal!)) AND exclusiv!`) required the term "sugar" and a term prefixed with "exclusiv" to be in the document for it to match. The vector query found 3 relevant documents in the first 294 retrieved that did not contain the term "sugar" (avt49c00, vie05d00 and zsc44c00) and 2 relevant documents which referred to "quota law" instead of "exclusivity" (gko97e00 and hko97e00). Note that the hko97e00 document counted as 8.6 relevant documents (because it's probability of being selected for judging was just 11.6%, and 1/0.116=8.6), which was almost half of the total estimated number of relevant documents (8.6/18.6=46%) (though the vector run would still have had a higher Recall@B for this topic than the boolean run even if the hko97e00 document was not counted). This observation is a reminder that the sampling-based scoring can be sensitive to a small number of judgments on individual topics.

2.3.5 Natural Language vs. Keywords

In this section, we refer to the vector of request text terms as the "natural language query" and the vector of the final boolean terms as the "keyword query". Note that these are both vector queries which do not use the boolean operators (besides the OR operator).

Using the natural language query led to a statistically significant increase in mean R@25000 compared to the using the keyword query (as per the "rvl-fv" entry of Table 3).

- Topic 87 ("All documents discussing Securities and Exchange Commission 10b-5 reports or reporting requirements"): This topic had the largest difference in R@25000 in favour of the natural language query (as per the "rvl-fv" entry of Table 3). For this query, the estimated number of relevant documents was 875. The Recall@25000 of the natural language query was 99.8%, while the keyword query's Recall@25000 was just 13%. The natural language query was (`'Securities'|'Exchange'|'Commission'|'10b'|'5'|'reports'|'reporting'|'requirements'`) while the keyword query was (`'10b'|'5'|'SEC'|'securities'|'exchange'|'commission'`). The key retrieval difference was that the natural language query retrieved the bgb65a00 document which counted for 758 relevant documents in the sampling procedure used. The reason for its relevance may have been that the following passage was interpreted as a reporting requirement: "Any person receiving this report and wishing to effect transactions in any security discussed herein should do so only with a representative of Morgan Stanley & Co. Incorporated or Dean Witter Reynolds Inc." The words in this sentence only matched the natural language query (in particular, the "report" concept was not in the keyword query, and "report" and "security" matched with the natural language query because it used inflection matching (from stemming)). The keyword query did have one concept not contained in the natural language query ("SEC"), but overall it appears that the natural language query had a higher recall from more concepts and the use of stemming.

Unfortunately we haven't had time to walk through more of the topic differences. We suspect in some cases the synonyms in the final negotiated query hurt precision over the first 25,000 retrieved (and hence hurt recall at that cutoff).

Note that compared to the final boolean query, the natural language query led to a statistically significant decline in mean R@B and P@B (as per the "rvl-fb" entries of Table 2).

2.3.6 Impact of Blind Feedback

In the past, we have found that our implementation of the blind feedback technique has been fairly reliable at boosting traditional IR measures, such as mean R-precision, including when using the final vector query as a baseline as in last year's experiments [12]. This year, we tried applying the blind feedback technique to the final boolean baseline to see if it would boost its Recall@B.

It turned out in this experiment that, unusually, the mean (raw) R-precision score fell with blind feedback (though the decline was not quite statistically significant as per the "fbe-fb" entry of Table 4). R@B and P@B also fell with blind feedback in this experiment (though again, these declines were not statistically significant as per the "fbe-fb" entries of Table 2).

The decline from blind feedback in the robust GS10J measure was statistically significant (as per the "fbe-fb" entry of Table 4) which is consistent with past experiments [10].

2.3.7 Impact of Fusion

Fusion runs often produce higher scores than any of their input runs, at least on the traditional IR metrics used in past experiments. In another attempt to increase the final boolean run's Recall@B, we tried a fusion run with the final boolean run as one of our inputs (50% weight), along with one-third weight on the request text run and one-sixth weight on the final vector run.

The mean (raw) R-precision showed a statistically significant increase with fusion compared to the final boolean run (as per the "frw-fb" entry of Table 4), so the fusion technique did seem to work on a traditional IR measure.

Table 5: Mean Scores of Submitted Relevance Feedback Task Runs

	Est.	Est.	Est.	Est.	Est.	Est.			(raw)
Run	R@B$_r$	R25000	P5	P@B$_r$	P25000	Gray@B$_r$	S1J	GS10J	R-Prec
otRF07fb	0.386	0.367	0.380	0.106	0.044	0.051	3/10	0.735	0.100
otRF07fv	0.248	0.444	0.355	0.156	0.064	0.007	3/10	0.612	0.065

Table 6: Impact of Dropping Boolean Operators on the 10 RF Task Topics

Measure	ΔMean	n/a	vs.	3 Extreme Diffs (Topic)
Est. R@B$_r$	−0.138	(−0.287, 0.012)	5-5-0	−0.54 (13), −0.47 (51), 0.14 (30)
Est. R25000	0.077	(−0.168, 0.321)	6-4-0	0.71 (30), −0.42 (37), −0.54 (13)
Est. P5	−0.025	(−0.236, 0.186)	1-3-6	0.75 (7), −0.20 (13), −0.60 (34)
Est. P@B$_r$	0.050	(−0.023, 0.122)	3-7-0	0.25 (34), 0.22 (7), −0.08 (37)
Est. P25000	0.020	(−0.037, 0.077)	6-4-0	0.17 (26), 0.15 (8), −0.15 (37)
Est. Gray@B$_r$	−0.044	(−0.084,−0.003)	1-5-4	−0.17 (34), −0.11 (26), 0.00 (45)
S1J	0.000	(−0.422, 0.422)	2-2-6	1.00 (8), 1.00 (7), −1.00 (27)
GS10J	−0.123	(−0.308, 0.061)	2-6-2	−0.69 (13), −0.36 (45), 0.37 (7)
(raw) R-Prec	−0.035	(−0.092, 0.021)	2-5-3	−0.21 (13), −0.13 (34), 0.11 (7)

However, the mean R@B and P@B scores both fell with fusion (though these declines were not statistically significant as per the "frw-fb" entries of Table 2). Table 2 shows that the biggest decline in R@B was on topic 58, which apparently suffered from fusing in the poor vector result described in section 2.3.4. In the raw R-Precision measure, though, the boolean run's score for topic 58 (0.15, or 6/41) was boosted by fusion (to 0.20, or 8/41).

Our fusion technique was fairly simple (it just added together the weighted rsv scores). But the rsv scores were not normalized. For example, in topic 58, the rsv scores ranged from 232 to 1 for the boolean run but just from 305 to 189 for the request vector run and just from 314 to 213 for the final vector run. Hence the fusion result tended to be dominated by the vector runs, especially in the deeper ranks, despite the higher weight (3x) on the boolean run (e.g. 3x a boolean rsv score of 1 would still be too small to have much impact compared to a bottom vector result of 189 or 213). We should perhaps investigate more sophisticated fusion schemes so that the deeper boolean results are not pushed out.

2.4 Relevance Feedback Task

The relevance feedback task re-used 10 topics from last year. Residual evaluation was used, i.e. documents judged last year were dropped from the results before evaluating.

We just submitted a couple of baseline runs that did not use relevance feedback. The submitted "otRF07fb" run used the same approach as the main task otL07fb final boolean run, and the submitted "otRF07fv" run used the same approach as the main task otL07fv final vector run. The mean scores of the runs are listed in Table 5; however, mean scores on 10 topics may not be so reliable. More insight may come from analyzing per-topic differences. Table 6 shows the largest per-topic differences in each measure when subtracting the boolean run's score from the vector run's score.

Based on an anecdotal look, the topics seem to be showing the same tendencies as last year regarding favoring the boolean or vector query. For example, last year [12] we listed topic 45 ("pigeon deaths") as one favoring the boolean run and topic 7 ("G-rated") as one favoring the vector run (based on various (raw) measures used last year). In this year's residual evaluation, we find the same tendency for these 2 topics; e.g. in this year's main measure, Est. R@B$_r$, the boolean query scores higher for topic 45 (0.48 vs. 0.70) while the vector query scores higher for topic 7 (0.08 vs. 0.01).

3 Glossary

3.1 Retrieval Measures

The retrieval measures of Tables 1 and 5 are defined as follows:

"Est. R@B" and "Est. R25000": Estimated Recall at Depths B and 25000.

"Est. P5", "Est. P@B" and "Est. P25000": Estimated Precision at Depths 5, B and 25000.

"Est. Gray@B": The estimated percentage of the first B documents that are Gray. ("Gray" documents were reviewed by the assessor but a judgement could not be made, such as because the document was longer than 300 pages, or there was a technical problem displaying the document.)

"S1J": Success of the First Judged Document. As only the first judged document is considered, the handling of duplicates does not affect this measure.

"GS10J": Generalized Success@10 on Judged Documents (1.08^{1-r} where r is the rank of the first relevant document, only counting judged documents, or zero if no relevant document is retrieved). GS10J is a robustness measure which exposes the downside of blind feedback techniques [10]. "Generalized Success@10" was originally introduced as "First Relevant Score" (FRS) in [11]. Intuitively, GS10J is a predictor of the percentage of topics for which a relevant document is returned in the first 10 rows.

"(raw) R-Prec": R-Precision (raw Precision at Depth R, where R is the raw number of known relevant documents). By "raw" we mean that estimation is not used for this measure, and unjudged documents are assumed non-relevant. This measure is our representative of "traditional" IR measures.

"B": number of documents matching the final negotiated boolean query.

"B_r" (RF task, residual evaluation): number of documents matching the final negotiated boolean query after removing matches that were judged last year.

3.2 Difference Tables

For the comparison tables (such as Table 2), the columns are as follows:

- "Expt" specifies the experiment (the codes of the two runs being compared are listed, indicating first run minus second run).
- "Δ" is the difference of the mean scores of the two runs being compared (the column heading says for which retrieval measure).
- "95% Conf" is an approximate 95% confidence interval for the mean difference (calculated from plus/minus twice the standard error of the mean difference). If zero is not in the interval, the result is "statistically significant" (at the 5% level), i.e. the feature is unlikely to be of neutral impact (on average), though if the average difference is small (e.g. <0.020) it may still be too minor to be considered "significant" in the magnitude sense.
- "vs." is the number of topics on which the first run scored higher, lower and tied (respectively) compared to the second run. These numbers should always add to the number of topics.
- "3 Extreme Diffs (Topic)" lists 3 of the individual topic differences, each followed by the topic number in brackets. The first difference is the largest one of any topic (based on the absolute value). The third difference is the largest difference in the other direction (so the first and third differences give the *range* of differences observed in this experiment). The middle difference is the largest of the remaining differences (based on the absolute value).

4 Conclusions

In our investigations of the boolean query negotiations, what we learned would probably make the plaintiff unhappy. The final negotiated boolean query averaged just 22% recall. Furthermore, the plaintiff's concessions to reduce the result set probably gave up more relevant documents than anticipated as the plaintiff boolean query averaged more than 32% recall (perhaps a lot more since our submission of some plaintiff queries were truncated at 25,000 results).

We found that doubling the proximity distances of the final boolean query would not have increased the average recall substantially (just from 22% to 24%), suggesting that the comparisons to the final boolean results likely are not overly sensitive to the exact proximity distances chosen by the negotiators.

We found that a couple of standard IR techniques, blind feedback and fusion, did not succeed at increasing Recall@B of the final boolean query (on average). Possibly an issue with our implementations was that the lower-ranked boolean results tended to be pushed out because they had relatively low rsv scores compared to the lower-ranked vector results.

We found that using the natural language request text as a source of vector terms produced a higher mean Recall@25000 than the using the keywords of the final boolean query as the vector source. Unfortunately we haven't yet investigated the reasons carefully enough to know if, say, the synonyms often included in the final boolean query tended to be harmful to the vector form of the query.

At depth-B, the final boolean query was the most successful at retrieving relevant documents of all the investigated techniques, on average. Our per-topic analysis suggested that the use of AND, phrase and proximity operators boosted precision at this depth. However, their use is partly responsible for the low average 22% recall of the final boolean query as we found that dropping the operators increased average recall to more than 35% (probably a lot more as our vector results were truncated at 25,000 documents).

That we can make statements about the relative success of techniques at depths B and 25000 is because of the sampling scheme used this year (in which each pooled document was judged with a known probability using a scheme we helped to design as part of organizing the track, as described in the track overview paper). The sparseness of the sampling is a concern; sometimes for individual topics we found that the retrieval of one relevant document (judged with low probability because no system retrieved it at a high rank) could have a substantial impact on the estimated score. Yet we also found that studying these high-weight documents (in particular, why one approach retrieved it and another did not) has been a helpful shortcut to finding plausible reasons for the success differences between approaches at substantial depths. Overall, we have gained many insights from analyzing the data, and our ability to gain even more insights into the reasons for the retrieval differences between techniques has been much more constrained by our limited time to walk through the individual topic results than by any lack of data.

References

[1] Jason R. Baron (Editor-in-Chief). The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery. The Sedona Conference Journal, Volume VIII, pp. 189-223, 2007.

[2] Jason R. Baron. Toward A Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery. The Sedona Conference Journal, Volume VI, pp. 237-246, 2005.

[3] Jason R. Baron, David D. Lewis and Douglas W. Oard. TREC-2006 Legal Track Overview. Proceedings of TREC 2006.

[4] Cross-Language Evaluation Forum web site. http://www.clef-campaign.org/

[5] Andrew Hodgson. Converting the Fulcrum Search Engine to Unicode. Sixteenth International Unicode Conference, 2000.

[6] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, J. Heard. Building a Test Collection for Complex Document Information Processing. *SIGIR 2006*, pp. 665-666.

[7] NTCIR (NII-Test Collection for IR) Home Page. http://research.nii.ac.jp/~ntcadm/index-en.html

[8] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford. Okapi at TREC-3. Proceedings of TREC-3, 1995.

[9] Text REtrieval Conference (TREC) Home Page. http://trec.nist.gov/

[10] Stephen Tomlinson. Early Precision Measures: Implications from the Downside of Blind Feedback. *SIGIR 2006*, pp. 705-706.

[11] Stephen Tomlinson. European Ad Hoc Retrieval Experiments with Hummingbird SearchServer$^{\mathrm{TM}}$ at CLEF 2005. Working Notes for the CLEF 2005 Workshop.

[12] Stephen Tomlinson. Experiments with the Negotiated Boolean Queries of the TREC 2006 Legal Discovery Track. Proceedings of TREC 2006.

[13] TREC 2007 Legal Discovery Track: Main Task Guidelines. http://trec-legal.umiacs.umd.edu/main07b.html