Experiments with the Negotiated Boolean Queries of the TREC 2006 Legal Discovery Track

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

February 8, 2007

Abstract

We analyze the results of several experimental runs submitted to the Legal Discovery and Terabyte Tracks of TREC 2006. In the Legal Track, the final negotiated boolean queries produced higher mean scores in average precision and Precision@10 than a corresponding vector run of the same query terms, but the vector run usually recalled more relevant items by rank 5000, and on average the boolean query matched less than half of the relevant items. We also report the impact of query negotiation, metadata and natural language vs. keywords. We find that robust metrics (which expose the downside of blind feedback techniques) are inappropriate for legal discovery because they favour duplicate filtering. We also report the results of depth probe experiments (depth 9000 in the Legal Track and depth 5000 in the Terabyte Track) which provide a lower-bound estimate for the number of unjudged relevant items in each track.

1 Introduction

Livelink ECM - eDOCS SearchServer™ (formerly known as Hummingbird SearchServer™) 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 [6] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (TREC [12], CLEF [4] and NTCIR [9]) 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 and terabyte retrieval.

2 Legal Discovery Track

The Legal Discovery Track (also known as the TREC Legal Track) was new to TREC this year. The collection to be searched was the IIT Complex Document Information Processing (IIT CDIP) test collection [7]. 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™ and SearchServer™ 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 (indeed, legal obligation) is to return all documents responsive (relevant) to a production request. For TREC, the organizers created 46 new production requests (topics), numbered from 6 to 51. 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), and a "final negotiated query" (a rejoinder boolean query from the plaintiff, which was used as the final query of the negotiations). (Examples appear in the topic analysis below.) During the assessing phase, 7 topics were dropped, leaving 39 topics. For these, 111 records were judged relevant per topic (low 1, high 502, median 63).

[19] has more details on the track and task, and [1] has more background on legal discovery in general.

2.1 Indexing

We indexed the collection twice, once including the full record (metadata plus ocr document), and a second time just indexing the ocr documents. For the full record case, 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. For the ocr-only case, we indexed from after the "<ot>" tag to the "</record>" tag, hence just the ocr document was in the FT_TEXT column. Any tags themselves were indexed (we just didn't bother to discard them). Entities (e.g. "&") were converted to the character they represented (e.g. "&").

For the full record index, 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.

For the ocr-only index, we used an English stopword list (e.g. "the", "of", "by") and punctuation was not indexed.

Both indexes just included the surface forms of the words (no 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 2006 (and 2 later re-runs) 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 [10] 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 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 (and 2 re-runs) were as follows:

humL06t (main boolean run): The submitted humL06t 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 45, for which the final negotiated query was "`(research OR stud! OR "in vivo") AND pigeon AND (death! OR dead OR die! OR dying)`", the corresponding SearchSQL statement was

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM LEGAL06FULL
WHERE ((FT_TEXT CONTAINS 'research', 'stud%', 'in vivo')
 AND    FT_TEXT CONTAINS 'pigeon'
 AND   (FT_TEXT CONTAINS 'death%', 'dead', 'die%', 'dying'))
ORDER BY REL DESC;
```

humL06t' (updated main boolean run (not submitted)): The humL06t' run was the same as humL06t except that an updated development build was used which fixed a proximity-matching bug that affected 7 of the queries. Fortunately, there was little impact on the overall results (more details are in Section 2.3.8).

humL06t0 (defendant boolean run): The submitted humL06t0 run was the same as humL06t except that the defendant query was used instead of the final negotiated boolean query. For example, for topic 45, for which the initial query proposed by the defendant was "`"animal studies" AND "pigeon deaths"`", the WHERE clause of the corresponding SearchSQL statement was

```
WHERE (FT_TEXT CONTAINS 'animal studies'
  AND  FT_TEXT CONTAINS 'pigeon deaths')
```

humL06t0' (updated defendant boolean run (not submitted)): The humL06t0' run was the same as humL06t0 except that the updated development build was used (more details are in Section 2.3.8).

humL06tv (main vector run): The submitted humL06tv run was the same as humL06t 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 45, the WHERE clause of the corresponding SearchSQL statement was

```
WHERE ((FT_TEXT CONTAINS 'research'|'stud%'|'in'|'vivo'|
                'pigeon'|'death%'|'dead'|'die%'|'dying'))
```

humL06tvo (main ocr-only run): The submitted humL06tvo run was the same as humL06tv except that the ocr-only index was used instead of the full-record index (i.e. the FROM clause of the SearchSQL statement referred to the LEGAL06OCR table instead of LEGAL06FULL).

humL06dvo (main natural language run): The submitted humL06dvo run was the same as humL06tvo except that the terms were taken from the request text instead of the final negotiated query. Terms in our list of common instruction words (e.g. "find", "relevant", "document") were discarded (using the same list we used in past years for the Description field of an ad hoc topic). For example, for topic 45, for which the request text was "All documents that refer or relate to pigeon deaths during the course of animal studies.", the WHERE clause of the corresponding SearchSQL statement was

```
WHERE FT_TEXT CONTAINS 'All'|'that'|'refer'|'or'|'relate'|'to'|
  'pigeon'|'deaths'|'during'|'the'|'course'|'of'|'animal'|'studies'
```

humL06tve (blind feedback run): The submitted humL06tve run was a blind feedback run based 50% on humL06tv and 25% each on expansion queries from the first 2 rows of humL06tv.

humL06tvc (duplicate filtering run): The submitted humL06tvc run was the same as humL06tv except that rows which appeared to be duplicates of a previous row were discarded. The heuristic filtering approach was to discard a row if all of the passages surrounding matching terms were identical to the set of such passages of any previous row.

humL06tvz (depth probe run): The first 90 rows of the submitted humL06tvz run were a one percent subset of the first 9000 rows of humL06tv (rows 1, 101, 201, 301, ..., 8901), and its remaining 1000 rows were rows 9001-10000 of humL06tv. (This run was based on the same approach as the humT06xlz run submitted a month earlier for the Terabyte track, described in a later section.)

For each run, only 5000 rows were allowed to be submitted for each query.

2.3 Results

Tables 1-3 list several mean scores for the 8 submitted runs (and 2 re-runs in parentheses). The retrieval measures are defined in Section 4.1 of the Glossary at the end of the paper. The highest mean scores of each measure are in bold; however, see Tables 4-10 for which mean differences are statistically significant. (The columns of Tables 4-10 are explained in Section 4.2 of the Glossary.)

Table 1: Mean Scores of Legal Discovery Runs (All 39 Topics)

Run	GS10	S10	S1	P10	R-Prec	bpref	GMAP	MAP	R@5000
(humL06t')	0.620	27/39	13/39	0.333	0.159	0.199	0.026	0.110	0.481
humL06t	0.615	26/39	13/39	**0.338**	**0.161**	**0.203**	**0.026**	**0.111**	0.502
humL06tve	0.534	21/39	12/39	0.310	0.135	0.198	0.020	0.097	0.612
humL06tv	0.609	27/39	12/39	0.277	0.119	0.170	0.019	0.080	**0.623**
humL06tvc	0.611	27/39	12/39	0.274	0.119	0.169	0.019	0.079	0.614
humL06dvo	0.587	25/39	9/39	0.303	0.103	0.189	0.014	0.065	0.483
humL06tvo	**0.627**	27/39	13/39	0.274	0.119	0.173	0.015	0.078	0.547
humL06t0	0.527	22/39	11/39	0.297	0.080	0.089	0.001	0.059	0.146
(humL06t0')	0.527	22/39	11/39	0.292	0.073	0.082	0.001	0.053	0.137
humL06tvz	0.494	20/39	12/39	0.113	0.036	0.029	0.001	0.011	0.057
[on judged]	GS10J	S10J	S1J	P10J	R-PrecJ	bprefJ	GMAPJ	MAPJ	R@5000J
(humL06t')	**0.644**	27/39	13/39	**0.344**	0.237	0.199	0.055	0.182	0.481
humL06t	0.634	26/39	13/39	0.338	0.241	**0.203**	0.056	**0.189**	0.502
humL06tve	0.557	21/39	12/39	0.310	**0.246**	0.198	**0.065**	0.186	0.612
humL06tv	0.614	27/39	12/39	0.277	0.212	0.170	0.058	0.165	**0.623**
humL06tvc	0.615	27/39	12/39	0.274	0.211	0.169	0.058	0.163	0.614
humL06dvo	0.619	25/39	9/39	0.310	0.233	0.189	0.056	0.157	0.483
humL06tvo	0.632	27/39	13/39	0.274	0.212	0.173	0.041	0.162	0.547
humL06t0	0.529	22/39	11/39	0.297	0.105	0.089	0.002	0.076	0.146
(humL06t0')	0.529	22/39	11/39	0.292	0.098	0.082	0.002	0.069	0.137
humL06tvz	0.494	20/39	12/39	0.118	0.038	0.029	0.002	0.012	0.057

Table 1 shows that the main vector run had a higher recall by depth 5000 (R@5000 measure) than the corresponding main boolean run, which is unsurprising because the vector run would usually match all of the same rows as the boolean run, and more.

However, if there was a penalty for retrieving non-relevant items before the relevant items, e.g. if precision was a factor in the measure, then the main boolean run typically had the higher mean score. In particular, the main boolean run had the highest mean average precision (MAP) of the listed runs of Table 1. MAP is a measure which requires both perfect recall and precision to get a 1.0 score.

Tables 2 and 3 include measures based on the boolean set size. In particular, Recall@B was less than 50% for the main boolean run, i.e. the boolean query did not match more than 50% of the (judged) relevant items (on average).

Note that only the first 10 rows submitted for each query were guaranteed to be in the assessing pool (and a few of those still were not assessed). Table 1 includes its measures after discarding unjudged items, and the results were similar (e.g. MAPJ gave almost the same ranking as MAP). Still, the inconsistency in the number of judged items in the boolean sets for different runs (J@B column of Table 2) is disconcerting.

The next 7 sections look at the differences between the submitted runs in more detail, followed by an additional section on the updated boolean runs.

2.3.1 Defendant Boolean vs. Main Boolean

The main boolean query (the rejoinder by the plaintiff) was apparently successful at finding more relevant items than the initial boolean query proposed by the defendant. Table 6 shows that the main boolean run's Recall@5000 (R@5000) was higher for 35 of the 39 topics and tied for the other 4, with no declines (as per the "t-t0" row for R@5000 of Table 6). We look at topic 45, for which Tables 4-7 show that the main boolean run was superior on several metrics, and topic 14, for which the defendant boolean run had a much higher Precision@10 (though still lower recall):

Table 2: Mean Scores of Legal Discovery Runs (23 Topics of B<5000)

Run	GS10	S1	P10	P@B	MAP	ret@B	J@B	PJ@B	R@B
(humL06t')	**0.695**	8/23	0.343	0.051	**0.139**	93%	38%	0.175	0.429
humL06t	0.679	8/23	0.343	0.052	0.137	100%	38%	0.167	**0.441**
humL06tve	0.519	8/23	0.274	0.056	0.098	100%	23%	0.225	0.373
humL06tv	0.613	7/23	0.248	0.050	0.084	100%	25%	0.202	0.354
humL06tvc	0.613	7/23	0.243	0.049	0.083	100%	25%	0.199	0.354
humL06dvo	0.629	4/23	**0.361**	0.038	0.084	100%	17%	**0.254**	0.333
humL06tvo	0.642	10/23	0.252	**0.091**	0.085	100%	25%	0.243	0.298
humL06t0	0.526	7/23	0.300	0.019	0.071	22%	42%	0.168	0.138
(humL06t0')	0.526	7/23	0.291	0.018	0.060	20%	45%	0.181	0.122
humL06tvz	0.570	7/23	0.109	0.005	0.010	78%	25%	0.049	0.046

Table 3: Mean Scores of Legal Discovery Runs (24 Topics of B2<5000)

Run	GS10	S1	P10	P@B2	MAP	ret@B2	J@B2	PJ@B2	R@B2
(humL06t')	**0.707**	9/24	0.363	0.057	0.135	100%	37%	0.186	**0.417**
humL06t	0.693	9/24	**0.371**	0.055	**0.136**	100%	37%	0.190	0.412
humL06tve	0.536	8/24	0.288	0.062	0.099	100%	24%	0.239	0.365
humL06tv	0.629	8/24	0.275	0.054	0.086	100%	26%	0.211	0.343
humL06tvc	0.629	8/24	0.271	0.053	0.085	100%	26%	0.208	0.342
humL06dvo	0.644	5/24	0.367	0.041	0.082	100%	19%	**0.259**	0.325
humL06tvo	0.654	10/24	0.275	**0.094**	0.085	100%	26%	0.250	0.292
humL06t0	0.543	7/24	0.296	0.020	0.068	22%	44%	0.169	0.132
(humL06t0')	0.543	7/24	0.288	0.019	0.057	20%	47%	0.182	0.118
humL06tvz	0.588	8/24	0.121	0.006	0.011	80%	27%	0.055	0.048

- Topic 45 ("pigeon deaths during the course of animal studies"): The defendant boolean query (`"animal studies" AND "pigeon deaths"`) required both the phrases "animal studies" and "pigeon deaths" to be in the record, which matched no records, and hence all of the relevant items were missed. The main boolean query (`(research OR stud! OR "in vivo") AND pigeon AND (death! OR dead OR die! OR dying)`) was more lenient, matching 2894 records, of which 387 were judged and 122 of those were judged relevant (although 35 other relevant records were not matched).
- Topic 14 ("marketing, placement, or sale of lozenges to children"): The defendant boolean query (`(marketing OR placement OR sale) AND "lozenges" AND "children"`) required the term "marketing" or "placement" or "sale" to be in the record, which led to high precision in the first 10 items (P10 of 0.9). The main boolean query (`lozenges AND (child! OR teen! OR juvenile! OR kid! OR adolescent)`) dropped the "marketing" requirement, which matched all the same documents and more (657 vs. 284) and found a couple more relevant items (19 instead of 17), but precision was lower (e.g. P10 was 0.0). The main boolean query still missed 17 of the 36 relevant items.

Overall, in the negotiations, it appears that the defendant typically proposed a high-precision query, and the plaintiff's rejoinder was to modify it so that it would match the same rows and more (higher recall), though sometimes at the expense of precision.

Table 4: Impact of Legal Discovery Techniques (GS10J, P10 and R-Prec measures)

Expt	ΔGS10J	95% Conf	vs.	3 Extreme Diffs (Topic)
t-t0	0.105	(−0.046, 0.256)	19-10-10	1.00 (44), 1.00 (45), −0.93 (51)
t'-t0'	0.115	(−0.036, 0.266)	19-10-10	1.00 (45), 1.00 (7), −0.93 (51)
t-tv	0.020	(−0.094, 0.133)	17-11-11	1.00 (17), −0.89 (39), −0.93 (33)
t'-tv	0.030	(−0.080, 0.139)	17-10-12	1.00 (17), −0.89 (39), −0.93 (33)
tv-tvo	−0.017	(−0.045, 0.010)	10-15-14	−0.27 (32), −0.25 (20), 0.26 (9)
dvo-tvo	−0.012	(−0.120, 0.096)	17-16-6	−0.92 (20), 0.71 (29), 0.86 (25)
tve-tv	−0.057	(−0.110,−0.005)	6-20-13	−0.63 (43), −0.55 (37), 0.23 (47)
tvc-tv	0.001	(−0.002, 0.004)	3-0-36	0.05 (37), 0.00 (24), 0.00 (51)
t'-t	0.010	(−0.006, 0.026)	3-1-35	0.22 (30), 0.21 (7), −0.07 (44)
t0'-t0	0.000	(−0.001, 0.001)	0-0-39	0.00 (29), 0.00 (7), 0.00 (51)
t-tve	0.077	(−0.049, 0.203)	18-11-10	1.00 (17), −0.89 (39), −0.93 (33)
t'-tve	0.087	(−0.035, 0.209)	18-10-11	1.00 (17), −0.89 (39), −0.93 (33)
	ΔP10			
t-t0	0.041	(−0.077, 0.159)	16-10-13	1.00 (45), 0.80 (21), −0.90 (14)
t'-t0'	0.041	(−0.069, 0.151)	16-10-13	1.00 (45), 0.70 (9), −0.90 (14)
t-tv	0.062	(−0.014, 0.138)	16-8-15	0.80 (45), 0.70 (32), −0.40 (18)
t'-tv	0.056	(−0.020, 0.133)	14-10-15	0.80 (45), 0.70 (32), −0.40 (18)
tv-tvo	0.003	(−0.018, 0.023)	5-4-30	−0.20 (35), 0.10 (13), 0.20 (9)
dvo-tvo	0.028	(−0.046, 0.102)	16-12-11	0.60 (45), 0.50 (30), −0.50 (13)
tve-tv	0.033	(−0.045, 0.112)	8-12-19	0.80 (45), 0.70 (27), −0.40 (26)
tvc-tv	−0.003	(−0.008, 0.003)	0-1-38	−0.10 (46), 0.00 (51), 0.00 (29)
t'-t	−0.005	(−0.022, 0.012)	2-2-35	−0.20 (7), −0.20 (21), 0.10 (44)
t0'-t0	−0.005	(−0.016, 0.006)	0-1-38	−0.20 (32), 0.00 (51), 0.00 (29)
t-tve	0.028	(−0.080, 0.136)	16-10-13	0.70 (32), −0.70 (27), −0.70 (39)
t'-tve	0.023	(−0.082, 0.128)	16-10-13	0.70 (32), −0.70 (27), −0.70 (39)
	ΔR-Prec			
t-t0	0.081	(0.039, 0.124)	25-4-10	0.50 (17), 0.29 (45), −0.17 (25)
t'-t0'	0.086	(0.043, 0.128)	25-3-11	0.50 (17), 0.32 (32), −0.17 (25)
t-tv	0.042	(−0.011, 0.096)	17-11-11	0.75 (17), 0.40 (32), −0.24 (33)
t'-tv	0.039	(−0.014, 0.093)	17-11-11	0.75 (17), 0.40 (32), −0.24 (33)
tv-tvo	0.000	(−0.005, 0.006)	11-8-20	0.06 (35), 0.04 (44), −0.03 (50)
dvo-tvo	−0.016	(−0.044, 0.012)	10-21-8	0.22 (39), −0.19 (9), −0.22 (37)
tve-tv	0.016	(−0.016, 0.047)	13-14-12	0.39 (39), 0.29 (45), −0.17 (32)
tvc-tv	−0.000	(−0.002, 0.001)	1-3-35	−0.02 (46), −0.00 (21), 0.01 (37)
t'-t	−0.003	(−0.012, 0.006)	1-5-33	0.12 (30), −0.06 (26), −0.09 (21)
t0'-t0	−0.007	(−0.021, 0.007)	0-1-38	−0.27 (32), 0.00 (51), 0.00 (29)
t-tve	0.027	(−0.040, 0.093)	14-16-9	0.75 (17), 0.57 (32), −0.44 (39)
t'-tve	0.024	(−0.042, 0.090)	13-16-10	0.75 (17), 0.57 (32), −0.44 (39)

2.3.2 Main Boolean vs. Main Vector

The main vector run used the same query terms as the main boolean run, but all boolean operators such as AND, phrases and proximities were dropped (all operators became an OR), and punctuation was dropped. Full wildcarding was still respected. The vector run always matched at least 5000 rows (the limit to submit).

Topics strongly favoring the boolean run (in at least one metric):

- Topic 45 ("pigeon deaths during the course of animal studies"): This topic had the

Table 5: Impact of Legal Discovery Techniques (bpref, GMAP' and MAP measures)

Expt	Δbpref	95% Conf	vs.	3 Extreme Diffs (Topic)
t-t0	0.114	(0.048, 0.181)	27-4-8	0.57 (45), 0.56 (17), −0.48 (51)
t'-t0'	0.118	(0.054, 0.182)	27-4-8	0.57 (45), 0.56 (17), −0.48 (51)
t-tv	0.032	(−0.031, 0.096)	19-12-8	0.75 (17), 0.39 (32), −0.39 (7)
t'-tv	0.029	(−0.037, 0.095)	19-12-8	0.75 (17), 0.39 (32), −0.40 (7)
tv-tvo	−0.003	(−0.011, 0.006)	10-19-10	−0.11 (32), −0.06 (9), 0.05 (21)
dvo-tvo	0.016	(−0.026, 0.059)	16-16-7	−0.42 (9), 0.32 (45), 0.33 (14)
tve-tv	0.027	(0.001, 0.053)	19-10-10	0.33 (39), 0.28 (45), −0.11 (37)
tvc-tv	−0.001	(−0.003, 0.001)	5-8-26	−0.02 (34), −0.02 (46), 0.01 (37)
t'-t	−0.003	(−0.023, 0.016)	3-3-33	−0.30 (21), −0.08 (26), 0.18 (30)
t0'-t0	−0.007	(−0.021, 0.007)	0-1-38	−0.27 (32), 0.00 (51), 0.00 (29)
t-tve	0.005	(−0.065, 0.076)	17-15-7	0.75 (17), 0.46 (32), −0.45 (9)
t'-tve	0.002	(−0.071, 0.075)	18-14-7	0.75 (17), 0.46 (32), −0.45 (9)
	ΔGMAP'			
t-t0	0.249	(0.144, 0.354)	29-7-3	0.89 (45), 0.79 (44), −0.27 (51)
t'-t0'	0.251	(0.147, 0.354)	31-5-3	0.89 (45), 0.79 (9), −0.27 (51)
t-tv	0.027	(−0.042, 0.096)	22-16-1	−0.82 (33), 0.52 (14), 0.59 (17)
t'-tv	0.027	(−0.043, 0.096)	23-15-1	−0.82 (33), 0.52 (14), 0.59 (17)
tv-tvo	0.021	(−0.003, 0.044)	18-17-4	0.37 (41), 0.21 (10), −0.04 (20)
dvo-tvo	−0.008	(−0.079, 0.062)	13-23-3	0.64 (41), 0.48 (14), −0.46 (20)
tve-tv	0.003	(−0.017, 0.024)	16-21-2	0.18 (45), 0.17 (39), −0.09 (44)
tvc-tv	−0.002	(−0.005, 0.001)	11-10-18	−0.04 (46), −0.03 (14), 0.00 (37)
t'-t	−0.000	(−0.007, 0.006)	3-3-33	−0.09 (21), 0.02 (20), 0.07 (30)
t0'-t0	−0.002	(−0.006, 0.002)	0-1-38	−0.07 (32), 0.00 (51), 0.00 (29)
t-tve	0.024	(−0.051, 0.098)	19-19-1	−0.84 (33), 0.56 (14), 0.62 (17)
t'-tve	0.023	(−0.051, 0.098)	19-19-1	−0.84 (33), 0.56 (14), 0.62 (17)
	ΔMAP			
t-t0	0.051	(0.007, 0.096)	29-7-3	0.63 (17), 0.29 (45), −0.29 (14)
t'-t0'	0.057	(0.013, 0.102)	31-5-3	0.63 (17), 0.29 (45), −0.29 (14)
t-tv	0.031	(−0.019, 0.081)	22-16-1	0.75 (17), 0.31 (32), −0.20 (31)
t'-tv	0.031	(−0.020, 0.081)	23-15-1	0.75 (17), 0.31 (32), −0.20 (31)
tv-tvo	0.001	(−0.003, 0.005)	18-17-4	0.03 (21), 0.03 (13), −0.03 (46)
dvo-tvo	−0.014	(−0.036, 0.008)	13-23-3	−0.20 (9), 0.14 (39), 0.18 (45)
tve-tv	0.017	(−0.006, 0.040)	16-21-2	0.32 (45), 0.26 (39), −0.05 (26)
tvc-tv	−0.001	(−0.002, 0.001)	11-10-18	−0.02 (46), −0.01 (34), 0.00 (37)
t'-t	−0.001	(−0.007, 0.006)	3-3-33	0.08 (30), −0.03 (26), −0.08 (21)
t0'-t0	−0.007	(−0.020, 0.007)	0-1-38	−0.26 (32), 0.00 (51), 0.00 (29)
t-tve	0.014	(−0.039, 0.068)	19-19-1	0.75 (17), 0.35 (32), −0.30 (39)
t'-tve	0.013	(−0.040, 0.067)	19-19-1	0.75 (17), 0.35 (32), −0.30 (39)

biggest difference in P10 in favour of the boolean run (as per the "t-tv" entry for ΔP10 of Table 4). The main boolean query (`((research OR stud! OR "in vivo") AND pigeon AND (death! OR dead OR die! OR dying)`) required the term "pigeon" to be in the record, which was good for precision (e.g. P10 of 1.0). In the main vector query (`'research'|'stud%'|'in'|'vivo'|'pigeon'|'death%'|'dead'|'die%'|'dying'`), the synonyms for "death" dominated the query, and many high-ranking matches did not mention pigeons, hurting precision (e.g. P10 of 0.2). At 5000 items retrieved (268 judged), the vector run had found 90 of the

Table 6: Impact of Legal Discovery Techniques (R@5000 measure)

Expt	ΔR@5000	95% Conf	vs.	3 Extreme Diffs (Topic)
t-t0	0.356	(0.274, 0.439)	35-0-4	1.00 (41), 0.80 (28), 0.00 (24)
t'-t0'	0.344	(0.265, 0.422)	35-0-4	1.00 (41), 0.80 (28), 0.00 (40)
t-tv	−0.120	(−0.226,−0.015)	11-24-4	−0.89 (24), −0.81 (33), 0.86 (51)
t'-tv	−0.142	(−0.250,−0.034)	11-24-4	−0.89 (24), −0.81 (33), 0.86 (51)
tv-tvo	0.075	(0.011, 0.140)	22-7-10	1.00 (41), 0.75 (17), −0.06 (46)
dvo-tvo	−0.064	(−0.167, 0.038)	7-27-5	1.00 (10), 1.00 (41), −0.56 (37)
tve-tv	−0.011	(−0.025, 0.003)	1-6-32	−0.20 (30), −0.17 (25), 0.01 (43)
tvc-tv	−0.008	(−0.018, 0.002)	5-10-24	−0.17 (47), −0.06 (34), 0.03 (20)
t'-t	−0.022	(−0.049, 0.006)	0-4-35	−0.50 (21), −0.19 (30), 0.00 (29)
t0'-t0	−0.009	(−0.027, 0.009)	0-1-38	−0.35 (32), 0.00 (51), 0.00 (29)
t-tve	−0.110	(−0.217,−0.003)	13-23-3	−0.89 (24), −0.81 (33), 0.86 (51)
t'-tve	−0.131	(−0.240,−0.022)	12-24-3	−0.89 (24), −0.81 (33), 0.86 (51)

Table 7: Impact of Legal Discovery Techniques (P@B and R@B, 23 Topics of B<5000)

Expt	ΔP@B	95% Conf	vs.	3 Extreme Diffs (Topic)
t-t0	0.033	(0.013, 0.054)	20-0-3	0.19 (31), 0.14 (46), 0.00 (33)
t'-t0'	0.033	(0.013, 0.054)	20-0-3	0.19 (31), 0.14 (46), 0.00 (33)
t-tv	0.002	(−0.013, 0.018)	11-7-5	−0.10 (7), 0.06 (46), 0.09 (43)
t'-tv	0.001	(−0.014, 0.017)	10-8-5	−0.10 (7), 0.06 (46), 0.09 (43)
tv-tvo	−0.042	(−0.129, 0.046)	9-4-10	−1.00 (33), 0.01 (35), 0.01 (31)
dvo-tvo	−0.053	(−0.140, 0.034)	5-16-2	−1.00 (33), −0.09 (31), 0.04 (14)
tve-tv	0.007	(−0.002, 0.015)	7-7-9	0.09 (46), 0.04 (35), −0.01 (43)
tvc-tv	−0.001	(−0.004, 0.002)	1-2-20	−0.03 (46), −0.00 (34), 0.00 (30)
t'-t	−0.001	(−0.003, 0.001)	0-2-21	−0.01 (30), −0.01 (26), 0.00 (33)
t0'-t0	−0.001	(−0.004, 0.002)	0-1-22	−0.03 (32), 0.00 (50), 0.00 (33)
t-tve	−0.004	(−0.021, 0.013)	11-8-4	−0.10 (7), −0.09 (35), 0.10 (43)
t'-tve	−0.005	(−0.022, 0.012)	10-9-4	−0.10 (7), −0.09 (35), 0.10 (43)
	ΔR@B			
t-t0	0.303	(0.191, 0.416)	20-0-3	1.00 (41), 0.78 (45), 0.00 (33)
t'-t0'	0.306	(0.201, 0.412)	20-0-3	1.00 (41), 0.78 (45), 0.00 (33)
t-tv	0.088	(−0.023, 0.198)	11-7-5	0.75 (17), 0.53 (14), −0.39 (7)
t'-tv	0.075	(−0.035, 0.185)	10-8-5	0.75 (17), 0.53 (14), −0.39 (7)
tv-tvo	0.056	(−0.032, 0.143)	9-4-10	1.00 (41), 0.14 (30), −0.03 (33)
dvo-tvo	0.036	(−0.099, 0.170)	5-16-2	1.00 (41), 0.67 (14), −0.31 (22)
tve-tv	0.020	(−0.009, 0.048)	7-7-9	0.21 (45), 0.20 (34), −0.07 (50)
tvc-tv	0.000	(−0.005, 0.005)	1-2-20	0.04 (30), −0.02 (34), −0.02 (46)
t'-t	−0.012	(−0.031, 0.006)	0-2-21	−0.19 (30), −0.09 (26), 0.00 (33)
t0'-t0	−0.015	(−0.046, 0.016)	0-1-22	−0.35 (32), 0.00 (50), 0.00 (33)
t-tve	0.068	(−0.050, 0.185)	11-8-4	0.75 (17), 0.54 (50), −0.41 (7)
t'-tve	0.055	(−0.061, 0.172)	10-9-4	0.75 (17), 0.54 (50), −0.41 (7)

157 relevant items, fewer than the boolean run which found 122 relevant items in the 2894 matched records (387 judged).

Table 8: Impact of Legal Discovery Techniques (PJ@B and J@B, 23 Topics of B<5000)

Expt	ΔPJ@B	95% Conf	vs.	3 Extreme Diffs (Topic)
t-t0	−0.000	(−0.072, 0.072)	12-8-3	−0.55 (34), 0.31 (43), 0.32 (45)
t'-t0'	−0.006	(−0.084, 0.072)	12-8-3	−0.55 (34), −0.44 (32), 0.32 (45)
t-tv	−0.034	(−0.087, 0.019)	8-12-3	−0.46 (7), −0.16 (34), 0.25 (43)
t'-tv	−0.027	(−0.083, 0.030)	8-12-3	−0.46 (7), 0.20 (30), 0.25 (43)
tv-tvo	−0.042	(−0.129, 0.046)	9-8-6	−1.00 (33), 0.03 (8), 0.04 (30)
dvo-tvo	0.011	(−0.094, 0.117)	13-8-2	−1.00 (33), 0.28 (34), 0.40 (14)
tve-tv	0.023	(0.003, 0.044)	11-6-6	0.17 (46), 0.10 (34), −0.03 (43)
tvc-tv	−0.002	(−0.008, 0.003)	4-3-16	−0.06 (46), −0.01 (34), 0.01 (30)
t'-t	0.007	(−0.010, 0.024)	3-1-19	0.19 (30), 0.00 (44), −0.02 (26)
t0'-t0	0.013	(−0.014, 0.040)	1-0-22	0.30 (32), 0.00 (8), 0.00 (50)
t-tve	−0.058	(−0.121, 0.006)	7-13-3	−0.49 (7), −0.26 (34), 0.28 (43)
t'-tve	−0.050	(−0.117, 0.017)	7-13-3	−0.49 (7), −0.26 (34), 0.28 (43)
	ΔJ@B			
t-t0	−0.041	(−0.260, 0.177)	8-14-1	1.00 (33), 1.00 (46), −0.92 (6)
t'-t0'	−0.069	(−0.290, 0.152)	8-14-1	1.00 (33), 1.00 (46), −0.92 (6)
t-tv	0.129	(0.068, 0.191)	20-2-1	0.57 (46), 0.37 (17), −0.02 (26)
t'-tv	0.134	(0.073, 0.195)	20-2-1	0.57 (46), 0.37 (17), −0.02 (26)
tv-tvo	0.002	(−0.002, 0.006)	12-6-5	0.03 (50), 0.02 (31), −0.01 (43)
dvo-tvo	−0.071	(−0.102,−0.040)	1-21-1	−0.35 (35), −0.18 (43), 0.01 (45)
tve-tv	−0.014	(−0.030, 0.002)	4-17-2	−0.12 (40), −0.08 (43), 0.06 (35)
tvc-tv	−0.001	(−0.004, 0.002)	2-5-16	−0.03 (46), −0.00 (34), 0.01 (30)
t'-t	0.005	(−0.012, 0.021)	2-3-18	0.17 (30), −0.01 (7), −0.06 (40)
t0'-t0	0.033	(−0.015, 0.080)	2-0-21	0.47 (32), 0.28 (41), 0.00 (50)
t-tve	0.143	(0.079, 0.207)	20-2-1	0.54 (46), 0.41 (17), −0.01 (26)
t'-tve	0.148	(0.085, 0.211)	20-2-1	0.54 (46), 0.41 (17), −0.02 (26)

- Topic 17 ("donations or contributions to the Libertarian Party"): This topic had the biggest difference in R@B, MAP, GMAP', bpref, R-Prec and GS10J in favour of the boolean run (as per the "t-tv" entries of Tables 4-7). The main boolean query (`((donat! OR contrib!) AND libertarian) BUT NOT (democrat! OR republic! OR GOP OR "G.O.P." OR "Grand Old Party")`), which excluded any records with words such as "Democratic", "Republican" or "GOP", matched 267 records (212 judged), finding 3 of the 4 relevant items (and the 3 relevant items were the first 3 items returned). The main vector query (`'donat%'|'contrib%'|'libertarian'|'democrat%'|'republic%'|'GOP'|'G'|'O'|'P'|'Grand'|'Old'|'Party'`) dropped the NOT operator, which hurt precision (the first relevant item was returned at rank 2478), though all 4 relevant items were returned in the first 5000 items (280 judged). (This was the only topic for which the main boolean query used a NOT operator.)

- Topic 43 ("contracts with medical supply companies or outfitters"): This topic had the biggest difference in P@B and PJ@B in favour of the boolean run (as per the "t-tv" entries of Tables 7 and 8). The main boolean query (`(contract! OR agreement! OR "purchase order" OR invoice) AND ("medical suppl!" OR outfitter!)`) matched 652 records (216 judged, 66 judged relevant). The main vector query (`'contract%'|'agreement%'|'purchase'|'order'|'invoice'|'medical'|'suppl%'|'outfitter%'`) found just 8 relevant items in the first 652 records. It appears that the "medical" concept was important for relevance, but the "suppl%" and "purchase" terms had more weight from inverse document frequency, and it hurt precision when these terms were not restricted to being

Table 9: Impact of Legal Discovery Techniques (P@B2 and R@B2, 24 Topics of B2<5000)

Expt	ΔP@B2	95% Conf	vs.	3 Extreme Diffs (Topic)
t-t0	0.035	(0.015, 0.056)	21-0-3	0.19 (31), 0.14 (46), 0.00 (33)
t'-t0'	0.038	(0.017, 0.059)	21-0-3	0.19 (31), 0.14 (46), 0.00 (33)
t-tv	0.000	(−0.014, 0.015)	10-9-5	−0.10 (7), 0.06 (46), 0.09 (43)
t'-tv	0.002	(−0.015, 0.019)	10-9-5	−0.10 (7), 0.08 (30), 0.09 (43)
tv-tvo	−0.039	(−0.123, 0.045)	10-4-10	−1.00 (33), 0.01 (31), 0.01 (21)
dvo-tvo	−0.053	(−0.136, 0.031)	5-17-2	−1.00 (33), −0.09 (31), 0.04 (14)
tve-tv	0.007	(−0.001, 0.016)	9-6-9	0.09 (46), 0.04 (35), −0.01 (43)
tvc-tv	−0.001	(−0.004, 0.002)	0-3-21	−0.03 (46), −0.00 (34), 0.00 (33)
t'-t	0.002	(−0.009, 0.013)	1-2-21	0.11 (30), −0.01 (26), −0.06 (21)
t0'-t0	−0.001	(−0.004, 0.002)	0-1-23	−0.03 (32), 0.00 (50), 0.00 (33)
t-tve	−0.007	(−0.023, 0.010)	10-10-4	−0.10 (7), −0.09 (35), 0.10 (43)
t'-tve	−0.005	(−0.024, 0.014)	11-9-4	−0.10 (7), −0.09 (35), 0.10 (43)
	ΔR@B2			
t-t0	0.280	(0.177, 0.383)	21-0-3	1.00 (41), 0.78 (45), 0.00 (33)
t'-t0'	0.299	(0.197, 0.401)	21-0-3	1.00 (41), 0.78 (45), 0.00 (33)
t-tv	0.069	(−0.037, 0.175)	10-9-5	0.75 (17), 0.53 (14), −0.39 (7)
t'-tv	0.073	(−0.037, 0.184)	10-9-5	0.75 (17), 0.53 (14), −0.39 (7)
tv-tvo	0.051	(−0.032, 0.135)	10-4-10	1.00 (41), 0.07 (6), −0.03 (33)
dvo-tvo	0.032	(−0.097, 0.161)	5-17-2	1.00 (41), 0.67 (14), −0.31 (22)
tve-tv	0.022	(−0.007, 0.050)	9-6-9	0.21 (45), 0.20 (34), −0.07 (50)
tvc-tv	−0.002	(−0.004, 0.001)	0-3-21	−0.02 (46), −0.02 (34), 0.00 (33)
t'-t	0.004	(−0.036, 0.045)	1-2-21	0.41 (30), −0.09 (26), −0.21 (21)
t0'-t0	−0.015	(−0.044, 0.015)	0-1-23	−0.35 (32), 0.00 (50), 0.00 (33)
t-tve	0.047	(−0.065, 0.160)	10-10-4	0.75 (17), 0.54 (50), −0.41 (7)
t'-tve	0.052	(−0.066, 0.169)	11-9-4	0.75 (17), 0.54 (50), −0.41 (7)

parts of phrases.

Topics strongly favoring the vector run (in at least one metric):

- Topic 7 ("company guidelines, strategies, or internal approval for placement of tobacco products in movies that are mentioned as G-rated"): This topic had the biggest differences in R@B, P@B, PJ@B and bpref in favour of the vector run (as per the "t-tv" entries of Tables 5-8). The main boolean query (`(((guide! OR strateg! OR approv!) AND (place! or promot!)) AND (("G-rated" OR "G rated" OR family) W/5 (movie! OR film! OR picture!))`) matched 645 records (226 judged, 9 relevant), while the vector query (`'guide%'|'strateg%'|'approv%'|'place%'|'promot%'|'G'|'rated'|'G'|'rated'|'family'|'movie%'|'film%'|'picture%'`) found 73 relevant items in the first 645. The boolean query formulation did not match relevant items with passages such as "`"G" and 'PG"-rated movies`" or even "`"G"-rated film`" (because the query only allowed 1 punctuation character between "G" and "rated"). Also, the assessor seems to have allowed "PG" as a synonym (e.g. "rated PG"). The boolean query's matches typically had "family" near a word like "moviegoer" or "film", which typically were not relevant. This topic is an example of the difficulties of forming a good boolean query.
- Topic 35 ("documents in which a tobacco company Chief Executive Officer or Chief Compliance Officer expressly refers to the Foreign Corrupt Practices Act"): This topic had the 2nd-biggest difference in Precision@B in favour of the vector run. The main boolean query (`(("Chief Executive Officer" OR "Chief Compliance Officer" OR CEO`

Table 10: Impact of Legal Discovery Techniques (PJ@B2 and J@B2, 24 Topics of B2<5000)

Expt	ΔPJ@B2	95% Conf	vs.	3 Extreme Diffs (Topic)
t-t0	0.021	(−0.054, 0.095)	13-8-3	−0.55 (34), 0.32 (45), 0.39 (21)
t'-t0'	0.005	(−0.073, 0.083)	13-8-3	−0.55 (34), −0.44 (32), 0.32 (45)
t-tv	−0.021	(−0.076, 0.034)	9-12-3	−0.46 (7), −0.16 (34), 0.25 (43)
t'-tv	−0.024	(−0.079, 0.030)	9-12-3	−0.46 (7), 0.22 (30), 0.25 (43)
tv-tvo	−0.039	(−0.123, 0.045)	10-8-6	−1.00 (33), 0.03 (8), 0.04 (21)
dvo-tvo	0.009	(−0.093, 0.111)	13-9-2	−1.00 (33), 0.28 (34), 0.40 (14)
tve-tv	0.028	(0.007, 0.049)	13-5-6	0.17 (46), 0.10 (21), −0.03 (43)
tvc-tv	−0.003	(−0.008, 0.003)	4-4-16	−0.06 (46), −0.01 (34), 0.00 (32)
t'-t	−0.003	(−0.018, 0.011)	2-3-19	−0.13 (21), −0.05 (26), 0.10 (30)
t0'-t0	0.013	(−0.013, 0.039)	1-0-23	0.30 (32), 0.00 (8), 0.00 (50)
t-tve	−0.049	(−0.112, 0.015)	8-13-3	−0.49 (7), −0.26 (34), 0.28 (43)
t'-tve	−0.052	(−0.116, 0.012)	7-14-3	−0.49 (7), −0.26 (34), 0.28 (43)
	ΔJ@B2			
t-t0	−0.076	(−0.296, 0.144)	8-15-1	1.00 (33), 1.00 (46), −0.92 (6)
t'-t0'	−0.104	(−0.328, 0.119)	8-15-1	1.00 (33), 1.00 (46), −0.92 (6)
t-tv	0.105	(0.038, 0.172)	19-4-1	0.57 (46), 0.37 (17), −0.25 (30)
t'-tv	0.108	(0.045, 0.171)	19-4-1	0.57 (46), 0.37 (17), −0.14 (21)
tv-tvo	0.002	(−0.003, 0.007)	13-6-5	−0.03 (40), 0.02 (31), 0.03 (50)
dvo-tvo	−0.073	(−0.107,−0.039)	2-21-1	−0.35 (35), −0.23 (30), 0.04 (40)
tve-tv	−0.017	(−0.034, 0.000)	4-17-3	−0.13 (40), −0.08 (43), 0.06 (35)
tvc-tv	−0.002	(−0.004, 0.001)	1-7-16	−0.03 (46), −0.00 (30), 0.00 (22)
t'-t	0.003	(−0.017, 0.023)	2-4-18	0.21 (30), −0.04 (40), −0.08 (21)
t0'-t0	0.031	(−0.014, 0.077)	2-0-22	0.47 (32), 0.28 (41), 0.00 (50)
t-tve	0.122	(0.055, 0.190)	19-4-1	0.54 (46), 0.41 (17), −0.17 (30)
t'-tve	0.125	(0.061, 0.190)	20-3-1	0.54 (46), 0.41 (17), −0.14 (21)

`OR "C.E.O." OR CCO OR "C.C.O.") AND ("Foreign Corrupt Practices Act" OR FCPA)`) matched 101 records, all of which were judged, with 9 judged relevant. The vector query (`'Chief'|'Executive'|'Officer'|'Chief'|'Compliance'|'Officer'|'CEO'|'C'|'E'|'O'|'CCO'|'C'|'C'|'O'|'Foreign'|'Corrupt'|'Practices'|'Act'|'FCPA'`) found 14 relevant items in the first 101 records. It appears that the key phrase for relevance was "Foreign Corrupt Practices Act" and that it was not necessary for a relevant item to contain the phrase "Chief Executive Officer" (or its synonyms in the boolean query). Some boolean matches were from stray terms such as "fCpA" and "ceo" appearing in a document, probably from ocr errors. Another relevant item (jup25f00) was missed by the boolean query because the ocr outputted "Chief Sxecutive o Officer".

It appears that the negotiations often led to a boolean query that was superior to a vector query of the same terms, but sometimes, perhaps from the lack of interaction with the documents, the negotiated query was less effective than it could have been. It is known from past TREC studies that good manual runs can outperform the top automatic runs. Perhaps the negotiated boolean queries can be thought of as "partial manual" queries in that they are manual but hindered from a lack of interactive feedback with the documents.

2.3.3 Impact of Metadata

The "tv-tvo" entry in Table 6 shows a statistically significant increase in R@5000, which intuitively makes sense (more relevant items are found when the metadata is not excluded).

The per-topic differences in MAP were small, however (no more than 0.03 on any topic, as per Table 5).

The mean difference for Precision@B was skewed by topic 33, for which B=1. Without metadata, a relevant item shifted from the 2nd to 1st rank, changing P@B from 0.0 to 1.0 for that topic.

2.3.4 Natural Language vs. Keywords

The "dvo-tvo" differences of Tables 4 and 6 suggest that using the natural language form of the query tended to have Precision@10, but lower Recall@5000, than using the keyword terms from the boolean query (though neither of these mean differences were statistically significant).

Topic 45 (on "pigeon deaths") may be illustrative. The synonyms for "death" in the boolean keyword list led to relatively less weight for "pigeon" in the vector query, hurting precision at the top of the list. But synonyms in the keyword list could produce higher recall (though for this topic, recall actually was still lower at 5000 items retrieved).

2.3.5 Impact of Blind Feedback

The "tve-tv" differences in Tables 4-8 show that the blind feedback technique caused a statistically significant decrease in GS10J, i.e. it pushed down the first relevant item (on average), but a statistically significant increase in PJ@B and bpref (and nearly significant increases in MAP and Precision@B). These results are consistent with what we have seen elsewhere [17, 13, 16, 15]. For example, in [15], 7 other groups' blind feedback systems (of the 2003 RIA workshop) were studied, and it was found that blind feedback was detrimental to the first relevant item (on average), even though it boosted the traditional TREC measures (such as P10, R-Prec and MAP).

Blind feedback is known to be bad for robustness because of its tendency to "not help (and frequently hurt) the worst performing topics" [21], hence most traditional TREC measures are non-robust, while measures of the first relevant item (such as S10 and GS10) appear to be robust.

[11] recently made the (unsupported) claim that for GMAP, "blind feedback is often found to be detrimental". However, in our past official experiments with GMAP ([16][14]) and in the RIA experiments ([15]) we have seen statistically significant increases in GMAP from blind feedback, but no statistically significant decreases. (In the case of our official Legal Track experiment, the GMAP measure was also slightly increased by blind feedback (from 0.019 to 0.020 as per Table 1) though this particular increase was not statistically significant.) We do not consider GMAP to be a robust measure.

An odd result is that there was a (nearly significant) decrease in R@5000 from blind feedback. We suspect this may be a case where the incompleteness of the assessments is producing a misleading result.

2.3.6 Impact of Duplicate Filtering

The "tvc-tv" differences in Table 4 found that duplicate filtering had the opposite impact to blind feedback, increasing robust measures (such as GS10J), and decreasing traditional TREC measures (such as P10, R-Prec and MAP) (though these particular results were not statistically significant, intuitively these results should hold up in larger experiments).

In legal discovery, one is obliged to turn over all responsive documents. One should not withhold items just because they are suspected of being duplicates. Hence "robust" measures are inappropriate for legal discovery.

In legal discovery, one should be consistent about what one returns, whereas robust metrics such as Success@10 encourage diversity [3].

(Duplicate filtering can be considered a special case of the incremental negative blind feedback approach derived for Success@10 in [3].)

Intuitively, the reason traditional TREC measures (such as P10, R-Prec and MAP) are not robust is that they encourage retrieval of duplicate information (and penalize duplicate filtering).

Intuitively, a recall-oriented measure would be robust if it just counted distinct aspects of a topic, such as the "instance recall" metric described in [3] (however, we presently have not done experiments with the

Table 11: Precision of Legal Discovery Run "humL06tv" at Various Depths

Depths	#Relevant (over 39 Topics)	Precision (Marginal)
1, 2, ..., 10	108 rel, 282 nonrel, 0 unjudged	0.277 (108/390)
101, 201, ..., 1001	37 rel, 350 nonrel, 3 unjudged	0.095 (37/390)
1101, 1201, ..., 2001	34 rel, 356 nonrel, 0 unjudged	0.087 (34/390)
2101, 2201, ..., 3001	32 rel, 356 nonrel, 2 unjudged	0.082 (32/390)
3101, 3201, ..., 4001	34 rel, 355 nonrel, 1 unjudged	0.087 (34/390)
4101, 4201, ..., 5001	27 rel, 362 nonrel, 1 unjudged	0.069 (27/390)
5101, 5201, ..., 6001	22 rel, 367 nonrel, 1 unjudged	0.056 (22/390)
6101, 6201, ..., 7001	17 rel, 370 nonrel, 3 unjudged	0.044 (17/390)
7101, 7201, ..., 8001	16 rel, 373 nonrel, 1 unjudged	0.041 (16/390)
8101, 8201, ..., 9001	17 rel, 371 nonrel, 2 unjudged	0.044 (17/390)
9002, 9003, ..., 9010	20 rel, 331 nonrel, 0 unjudged	0.057 (20/351)

"instance recall" measure, and such a metric requires special assessor effort to mark the distinct aspects (which was done for 20 topics for the TREC-6, 7 and 8 Interactive Tracks, according to [3])).

For most ad hoc tasks, we encourage the organizers to use a robust metric as the main measure, but legal discovery is an exception.

Intuitively, for non-feedback experiments (such as weighting experiments, or stemming, or boolean vs. vector), robust and non-robust metrics will tend to favor the same approaches. The differences between robust and non-robust metrics will be most evident for feedback experiments (robust metrics will favor negative feedback, such as duplicate filtering, and non-robust metrics will favor positive feedback, such as pseudo-relevance feedback (blind feedback)).

2.3.7 Depth Probe Results

Table 11 shows the (marginal) precision of the main vector run (humL06tv) at various depths, based on the first 10 rows of humL06tv (which were judged), and the first 100 rows of humL06tvz (rows 1, 101, 201, ..., 9001 of humL06tv, plus rows 9002-9010). humL06tvz was given highest judging precedence and most of its first 100 rows were judged.

As expected, precision tends to drop deeper in the result list, but even by depth 9000, the marginal precision was still approximately 4%.

For the 90 sample points of depths 101-9001 (x39 topics), Table 11 shows there were 236 relevant items (from 37+34+32+34+27+22+17+16+17), which projects to an estmate of 23600 relevant items in the first 9000 retrieved (over 39 topics), or 605 relevant items per topic. (This estimate is more likely to be lower than the true number of relevant items than higher, assuming precision tends to fall deeper in the result list, as our sample point was at the end of each 100 retrieved. Also, there are likely relevant items deeper than 9000, and there are likely relevant items that are not matched at all by the query.)

The actual number of judged relevant items was 111 per topic, hence apparently less than 20% of the relevant items are judged. Such incompleteness does not mean that the test collection is not useful; indeed, we have been able to learn a lot from the topic analysis above. But in general, one should be cautious about interpreting the mean scores without checking the individual topics.

2.3.8 Updated Boolean Runs

After the run submission deadline, a bug was found in the proximity matching of the development version used to produce the submitted boolean runs humL06t, humL06t0 and reference run humL06B (humL06B was the same as humL06t except that all matching rows were retrieved, not just the first 5000 rows).

Table 12: Topics Affected by the Boolean Update

Topic	B	incorrect	missed	B2
7	645	7	10	648
20	72113	3462	36	68687
21	9771	8683	3	1091
26	2968	405	0	2563
30	1359	1035	25	349
40	465	259	0	206
44	887	52	0	835

After the fix, 98% of the matches in humL06B were the same. 7 of the 46 queries were affected however (as listed in Table 12). In the original humL06B, 2% of the rows listed were incorrect, and 0.01% of the desired matches were missing.

(The issue was just in the development stream (not in released versions). On queries that requested a phrase within x words of a term, some new optimizations caused parts of the matching code to compare character positions to word positions, leading to incorrect results. When incorrect hits were produced, typically the document still satisfied the query if the proximity constraint was relaxed to a boolean-AND.)

When it was announced that 7 of the 46 topics had been dropped during the assessing phase, we originally assumed (incorrectly) that the dropped topics were the 7 affected by the bug, but in fact (as we verified after the conference) all 7 were still included in the final 39 topics.

Table 12 lists the 7 topics affected in the humL06B and humL06t runs. "B" is the original number of matching rows for the topic. "incorrect" is the number of original matching rows that should not have matched. "missed" is the number of desired rows missed by the original humL06B run. "B2" is the number of matches in the updated boolean run.

In every table reporting mean scores (i.e. Tables 1-3), we have listed both the original submitted boolean runs (humL06t and humL06t0) and the updated versions of these runs (humL06t' and humL06t0'). The mean scores of the original and updated runs are very similar.

For every measure based on the first B items retrieved (i.e. the @B measures of Table 2), we have also listed the scores based on the first B2 items retrieved (in Table 3). The mean scores of the @B and @B2 measures are very similar.

In the difference tables (Tables 4-10), for every comparison involving the boolean runs (t and t0), such as "t-tv", we have also listed the comparison using the updated boolean runs (t' and t0'), e.g. "t'-tv". Again, the results are very similar, and in particular the extreme topics are usually the same.

Furthermore, for every measure in Tables 4-10, we have also added direct comparisons of the updated and original runs (as per the "t'-t" and "t0'-t0" entries of these tables). For instance, for the R@B2 and P@B2 measures (Table 9), we see that just 3 topics were affected in the "t'-t" case (topics 30, 21 and 26), and the largest difference was in the opposite direction to the other two differences (mostly cancelling their impact on the mean difference).

Of the topics walked through earlier, the only one affected by the update was topic 7, which was the one least affected. The original analysis for this topic remains valid.

Fortunately, it appears that whether one uses the original or updated boolean run for analysis is unlikely to have much impact on one's conclusions.

2.4 Comparison to Previous Boolean Studies

[20] reported that automatic ranking of natural language queries substantially outperformed manually produced Boolean queries in an experiment, which sounds quite different from our result. [20] used date-ordering on the Boolean set instead of relevance-ranking, which might have disadvantaged the Boolean run on measures looking less than B items deep. Furthermore, many of the metrics used by [20] looked deeper than the

Table 13: Mean Scores of Submitted Terabyte Adhoc Runs

Run	GS30	GS10	S10	MRR	S1	P10	bpref	GMAP	MAP
humT06xlc	0.964	0.932	49/50	0.796	35/50	0.630	0.361	0.203	0.298
humT06xl	0.963	0.931	49/50	0.793	35/50	0.632	0.387	0.222	0.327
(humT06x5l)	0.956	0.917	48/50	0.808	37/50	0.588	0.404	0.233	0.327
humT06l	0.956	0.912	48/50	0.777	35/50	0.562	0.364	0.188	0.296
humT06xle	0.947	0.899	46/50	0.773	35/50	0.654	0.418	0.232	0.345
(humT06)	0.937	0.887	47/50	0.753	34/50	0.558	0.347	0.161	0.274
humT06xlz	0.827	0.797	42/50	0.732	35/50	0.198	0.026	0.006	0.016

median boolean result set size (e.g. the median was B=16 in the main experiment). [20] does not report the Precision@B and Recall@B measures, arguing that measures based on the boolean set size are potentially biased in favor of the Boolean system.

[8] reported that manually produced Boolean queries performed similarly to the top-performing automatic and manual rankings of the TREC-4 topics submitted to TREC-4, which sounds similar to our result. [8] used the Precision@B and Recall@B measures. The B values ranged from 5 to 288 (whereas ours ranged from 1 to 4183, excluding topics with B>5000).

[20] and [8] both produced their own boolean queries, whereas we used the negotiated boolean queries of the Legal Track topic set.

We performed experiments isolating the boolean operators (i.e. vector runs based on the same keywords as the boolean query), whereas [20] and [8] did not include such an experiment. The experiments in [20] and [8] seem more analogous to comparisons of the request text and the boolean query, which did not always use the same words.

For our main experiments, we identified and walked through the extreme topics (those with the greatest differences in each direction in various metrics) to help understand and verify the reasons for the differences. [20] and [8] did not include such topic analysis.

3 Terabyte Track

For the tasks of the Terabyte Track, the collection to be searched was the GOV2 collection, a crawl of most of the .gov domain in early 2004. Once binaries (such as images) were removed, its size was less than half a terabyte. The GOV2 distribution was 457,165,206,582 bytes uncompressed (426 GB) and consisted of 25,205,179 documents. More than 90% of the documents were html, 8% were (extracted text from) pdf, and the rest were extracted text from other formats (plain text, msword, postscript, etc.). The average document size was 18,137 bytes.

We participated in all 3 tasks of the Terabyte Track: adhoc, efficiency and named page finding. Details on these tasks are in the track guidelines [18].

3.1 Adhoc Experiments

In the Adhoc Task of the Terabyte Track, there were 50 topics, each with a Title, Description and Narrative field. The terabyte adhoc judgements contained on average 118 relevant documents per topic (low 6, high 571, median 87.5) counting both "relevant" and "highly relevant" as relevant.

The techniques used for the 5 submitted runs of June 2006 (and 2 other diagnostic runs done at the same time):

humT06 (not submitted): The SearchServer CONTAINS predicate was used to perform a boolean-OR of the words of the Title field of the topic. (No inflections.)

humT06l: Same as humT06 except that it included linguistic expansion from English inflectional stemming.

Table 14: Impact of Terabyte Adhoc Techniques

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
l (l-none)	0.026	(−0.017, 0.069)	8-7-35	0.79 (830), 0.49 (840), −0.32 (806)
x (xl-l)	0.018	(−0.002, 0.039)	8-2-40	0.32 (806), 0.27 (805), −0.13 (834)
x5 (x5l-l)	0.004	(−0.039, 0.048)	9-4-37	−0.73 (827), −0.46 (829), 0.32 (806)
c (xlc-xl)	0.001	(−0.002, 0.005)	1-1-48	0.07 (804), 0.00 (802), −0.00 (803)
e (xle-xl)	−0.032	(−0.071, 0.008)	4-11-35	−0.75 (830), −0.39 (805), 0.32 (825)
	ΔMAP			
x (xl-l)	0.032	(0.020, 0.044)	40-8-2	0.13 (831), 0.13 (821), −0.05 (827)
x5 (x5l-l)	0.032	(0.000, 0.063)	25-24-1	0.36 (806), 0.29 (849), −0.16 (827)
l (l-none)	0.021	(−0.001, 0.043)	27-19-4	0.18 (839), 0.18 (802), −0.15 (826)
e (xle-xl)	0.018	(−0.009, 0.045)	29-20-1	0.32 (847), 0.18 (831), −0.31 (816)
c (xlc-xl)	−0.029	(−0.041,−0.016)	8-40-2	−0.18 (819), −0.13 (808), 0.08 (818)

humT06xl: Same as humT06l except that a small additional weight (20%) was put on matching all of the query words within 200 characters of each other (ignoring stopwords which were not indexed).

humT06x5l (not submitted): Same as humT06xl except that the weight was 5-to-1 in favour of the proximity predicate instead of the boolean-OR. (So this run would be likely to rank all documents which matched the proximity predicate ahead of those that did not.)

humT06xle: Blind feedback run based 50% on humT06xl and 25% each on expansion queries from the first 2 rows of humT06xl.

humT06xlc: Same as humT06xl except that rows which appeared to be duplicates of a previous row were discarded. The heuristic filtering approach was was to discard a row if all of the passages surrounding matching terms were identical to the set of such passages of any previous row.

humT06xlz (depth probe run): The first 90 rows of the submitted humT06xlz run were a one percent subset of the first 9000 rows of humT06xl (rows 1, 101, 201, 301, ..., 8901), and its remaining 1000 rows were rows 9001-10000 of humT06xl. (The organizers requested runs to try to find "unique" relevant items.)

For each run, only 10000 rows were allowed to be submitted for each query.

Table 13 lists the mean scores of the adhoc runs, and Table 14 isolates the differences between the runs. The results were similar to last year's [16]; in particular, blind feedback had the anticipated effect of boosting the non-robust MAP measure but decreasing the robust GS10 measure. The main new experiment this year was duplicate filtering, which modestly increased the robust GS10 measure, while causing a statistically significant decrease on the non-robust MAP measure. (These feedback results are consistent with what we saw in our Legal Track experiments.)

Table 15 shows the (marginal) precision of the humT06xl run at various depths, based on the first 50 rows of humT06xl and the first 50 rows of humT06xlz.

If we do a similar calculation to what we described for the Legal Track, then we produce an estimate of 16200 relevant items in the first 4900 rows (from (67+30+31+20+14)*100), or 324 relevant items per topic. The actual number of judged relevant items was 118 per topic, hence apparently less than 40% of the relevant items are judged. (Again, this does not mean that the test collection is not useful, just that one should be cautious.)

Table 16 has a more detailed breakdown of the number and type of items retrieved at each depth. The "relevant" items of Table 15 are separated into "highly relevant" and "other relevant" in Table 16; we see that highly relevant items are particularly prevalent in the first 10 ranks. The "non-relevant" items of Table 15 are separated into "(judged) non-relevant" and "unretrieved" in Table 16 (the latter occurred because in some cases the Title query matched fewer than 4901 rows). Table 16 also includes items retrieved at depths 51-100, from which we can see the number of unjudged relevant items in an unpooled area of the humT06xl run (e.g. the marginal rate of unjudged items exceeds 30% at depths 91-100). The 3rd column of Table 16

Table 15: Precision of Terabyte Adhoc Run "humT06xl" at Various Depths

Depths	#Relevant (over 50 Topics)	Precision (Marginal)
1, 2, ..., 10	316 rel, 184 nonrel, 0 unjudged	0.632 (316/500)
11, 12, ..., 20	225 rel, 275 nonrel, 0 unjudged	0.450 (225/500)
21, 22, ..., 30	207 rel, 293 nonrel, 0 unjudged	0.414 (207/500)
31, 32, ..., 40	207 rel, 293 nonrel, 0 unjudged	0.414 (207/500)
41, 42, ..., 50	181 rel, 319 nonrel, 0 unjudged	0.362 (181/500)
101, 201, ..., 1001	67 rel, 433 nonrel, 0 unjudged	0.134 (67/500)
1101, 1201, ..., 2001	30 rel, 470 nonrel, 0 unjudged	0.060 (30/500)
2101, 2201, ..., 3001	31 rel, 469 nonrel, 0 unjudged	0.062 (31/500)
3101, 3201, ..., 4001	20 rel, 480 nonrel, 0 unjudged	0.040 (20/500)
4101, 4201, ..., 4901	14 rel, 436 nonrel, 0 unjudged	0.031 (14/450)

Table 16: More Detailed Breakdown of Terabyte Adhoc Run "humT06xl" at Various Depths

Depths	Breakdown of Retrieved Items (over 50 Topics)	#H/#R
1, 2, ..., 10	71 highly rel, 245 other rel, 184 nonrel, 0 unret, 0 unjudged	22%
11, 12, ..., 20	17 highly rel, 208 other rel, 275 nonrel, 0 unret, 0 unjudged	8%
21, 22, ..., 30	22 highly rel, 185 other rel, 293 nonrel, 0 unret, 0 unjudged	11%
31, 32, ..., 40	18 highly rel, 189 other rel, 293 nonrel, 0 unret, 0 unjudged	9%
41, 42, ..., 50	12 highly rel, 169 other rel, 319 nonrel, 0 unret, 0 unjudged	7%
51, 52, ..., 60	15 highly rel, 149 other rel, 271 nonrel, 0 unret, 65 unjudged	9%
61, 62, ..., 70	9 highly rel, 128 other rel, 272 nonrel, 0 unret, 91 unjudged	7%
71, 72, ..., 80	13 highly rel, 121 other rel, 242 nonrel, 0 unret, 124 unjudged	10%
81, 82, ..., 90	11 highly rel, 117 other rel, 235 nonrel, 0 unret, 137 unjudged	9%
91, 92, ..., 100	5 highly rel, 98 other rel, 228 nonrel, 0 unret, 169 unjudged	5%
101, 201, ..., 1001	2 highly rel, 65 other rel, 433 nonrel, 0 unret, 0 unjudged	3%
1101, 1201, ..., 2001	0 highly rel, 30 other rel, 470 nonrel, 0 unret, 0 unjudged	0%
2101, 2201, ..., 3001	1 highly rel, 30 other rel, 467 nonrel, 2 unret, 0 unjudged	3%
3101, 3201, ..., 4001	1 highly rel, 19 other rel, 470 nonrel, 10 unret, 0 unjudged	5%
4101, 4201, ..., 4901	0 highly rel, 14 other rel, 427 nonrel, 9 unret, 0 unjudged	0%

lists the percentage of all relevant items of the depth range which were judged highly relevant; it appears that this percentage is not a constant, but decreases with retrieval depth.

3.2 Efficiency Experiments

For the efficiency task, the 50 adhoc topics and 181 named page topics were seeded into a set of 100,000 queries to run. Runs were due before the adhoc and named page topics were identified. We submitted two runs.

The humTE06i3 run was like the ad-hoc humT06 run except that it used boolean-AND instead of boolean-OR and did not enable document length normalization. It averaged 1.7 seconds per query on our 2.8GHz machine. On the ad-hoc topics, it had an MRR of 0.725 and P10 of 0.416, and on the named-paged topics (below), it had an MRR of 0.123 and S10 of 40/181.

The humTE06v2 run was like the named-page humTN06pl run (below) except that the phrase matching on the title fields was omitted (though vector matching on the title fields was still included) and query terms in more than 10% of the rows were discarded. It averaged 4.6 seconds per query. On the ad-hoc topics, it

Table 17: Mean Scores of Submitted Terabyte Named Page Finding Runs

Run	GS10	S1	S5	S10	S1000	MRR
(humTN06dp)	0.568	61/181	95/181	105/181	157/181	0.425
humTN06dpl	0.567	55/181	98/181	103/181	157/181	0.408
humTN06dplc	0.565	55/181	97/181	104/181	155/181	0.406
humTN06pl	0.564	56/181	94/181	110/181	152/181	0.407
(humTN06rdpl)	0.491	39/181	80/181	92/181	159/181	0.319
humTN06l	0.414	32/181	62/181	79/181	144/181	0.262

Table 18: Impact of Terabyte Named Page Finding Techniques

Expt	ΔGS10	95% Conf	vs.	3 Extreme Diffs (Topic)
p (pl-l)	0.150	(0.102, 0.198)	94-32-55	1.00 (1049), 1.00 (1055), −0.61 (1006)
d (dpl-pl)	0.003	(−0.023, 0.028)	35-48-98	0.93 (999), 0.74 (1018), −0.73 (1080)
l (dpl-dp)	−0.001	(−0.026, 0.024)	28-46-107	1.00 (1021), 0.93 (1008), −0.61 (958)
c (dplc-dpl)	−0.002	(−0.013, 0.009)	28-4-149	−0.93 (902), 0.12 (959), 0.12 (1063)
r (rdpl-dpl)	−0.076	(−0.112,−0.040)	19-89-73	1.00 (1048), −0.93 (1074), −1.00 (1023)

had an MRR of 0.588 and P10 of 0.454. On the named-page topics, it had an MRR of 0.373 and S10 of 95/181.

3.3 Named Page Finding Experiments

For the 181 queries of the Named Page Finding Task, the goal was to find a particular named page. Table 17 lists the mean scores of the 4 submitted runs and 2 other runs saved at the same time, and Table 18 isolates the differences between the runs. Most of the run approaches and results are similar to last year's [16].

The main new experiment was the 'c' experiment (duplicate filtering heuristic): Table 18 shows that it improved the result for 28 queries and only hurt 4 queries, though one of the hurt queries (topic NP902 (home reference definition of embryo)) had a substantial drop in score that led to a negative mean difference in GS10. However, the result for NP902 was because of the assessments not identifying GX016-03-2878286 (http://ghr.nlm.nih.gov/ghr/glossary/embryo;jsessionid=3BE5FDA29D238576F653BFEE4DEB1BC1) as a duplicate of the list of accepted pages (such as GX018-73-5535473 (http://ghr.nlm.nih.gov/ghr/glossary/embryo;jsessionid=F794F12416497D73D1C185471CE91BE9)).

4 Glossary

4.1 Retrieval Measures

The ad hoc retrieval measures of Tables 1 and 13 are defined as follows:

- *Recall@5000* (R@5000): For a topic, R@5000 is the percentage of the relevant items retrieved in the first 5000 rows.
- *Precision@n*: For a topic, "precision" is the percentage of retrieved documents which are relevant. "Precision@n" is the precision after n documents have been retrieved. This paper lists Precision@10 (P10) for all ad hoc runs.

- *R-Precision*: For a topic, R-Prec is the precision at rank R, where R is the number of relevant items for the topic.
- *Average Precision* (AP): For a topic, AP is the average of the precision after each relevant item is retrieved (using zero as the precision for relevant items which are not retrieved). (In this paper, AP is based on the first 5000 retrieved items for the Legal runs and the first 10000 retrieved items for Terabyte runs.) The score ranges from 0.0 (no relevants found) to 1.0 (all relevants found at the top of the list). "Mean Average Precision" (MAP) is the mean of the average precision scores over all of the topics (i.e. all topics are weighted equally).
- *Binary Preference*: bpref is based just on the relative ranking of judged items, defined in [2].
- *Geometric MAP* (GMAP): GMAP (introduced in [22]) is based on "Log Average Precision" which for a topic is the natural log of the max of 0.00001 and the average precision. GMAP is the exponential of the mean log average precision.
- *GMAP'*: GMAP' linearly maps the 'log average precision' values to the [0,1] interval, putting the individual topic scores on the same scale as the other measures and allowing the arithmetic mean to be used as normal. GMAP' produces the same system rankings as GMAP and the same conclusions for statistical significance purposes.
- *Reciprocal Rank* (RR): For a topic, RR is $\frac{1}{r}$ where r is the rank of the first relevant item, or zero if no relevant item is retrieved. "Mean Reciprocal Rank" (MRR) is the mean of the reciprocal ranks over all the topics.
- *Success@n* (S@n): For a topic, Success@n is 1 if a desired page is found in the first n rows, 0 otherwise. This paper lists Success@10 (S10) and Success@1 (S1) for most runs. (Success@5 (S5) and Success@1000 (S1000) are also sometimes listed.)
- *Generalized Success@10* (GenS@10 or GS10): For a topic, GS10 is 1.08^{1-r} where r is the rank of the first relevant item, or zero if no relevant item is retrieved. (This measure was known as "First Relevant Score" (FRS) last year.) GS10 is considered a generalization of Success@10 because it rounds to 1 for $r \leq 10$ and to 0 for $r > 10$.
- *Generalized Success@30* (GenS@30 or GS30): For a topic, GS30 is 1.024^{1-r} where r is the rank of the first relevant item, or zero if no relevant item is retrieved.

We attach a J suffix to the measure (e.g. GS10J, S10J, S1J, MAPJ, P10J, R-PrecJ, bprefJ, GMAPJ, MAPJ, R@5000J) when unjudged items are omitted rather than being assumed non-relevant. Note that our MAPJ measure is the same as the "induced AP (indAP)" measure of [24].

For all measures, the mean scores weight each included topic equally.

The boolean set measures of Tables 2 and 3 are defined as follows:

- *B*: B is the number of items matched by the main boolean run (humL06t). For the @B measures, topics with B≥5000 are omitted (leaving just 23 topics instead of the usual 39 for Legal runs).
- *Precision@B*: For a topic, Precision@B (P@B) is the precision after B items have been retrieved. (If fewer than B items were retrieved, the number of relevant items retrieved is still divided by B.)
- *PJ@B*: For a topic, PJ@B is the same as P@B except that the precision is just based on the judged items in the first B retrieved; if no judged items were retrieved in the first B items, then a score of 0 was assigned.
- *Recall@B*: For a topic, Recall@B (R@B) is the percentage of the relevant items retrieved in the first B rows.

- *Judged@B*: For a topic, Judged@B (J@B) is the percentage of the first B retrieved items that are judged (relevant or non-relevant). If fewer than B items were retrieved, we (perhaps unfortunately) divided the number of judged items by the actual number of retrieved items (not B).
- *Retrieved@B*: For a topic, Retrieved@B (ret@B) is the number of retrieved items divided by B (capped at 100% for each topic).
- *B2*: B2 is the number of items matched by the updated boolean run (humL06t'). P@B2, PJ@B2, R@B2, J@B2 and ret@B2 are the same as P@B, PJ@B, R@B, J@B and ret@B (respectively) except that the first B2 retrieved items are examined instead of the first B. For the @B2 measures, topics with B2≥5000 are omitted (leaving just 24 topics).

4.2 Difference Tables

For the comparison tables (Tables 4-10, 14 and 18), the columns are as follows:

- "Expt" specifies the experiment (the codes of the two runs being compared are in parentheses, 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).

5 Conclusions

In the Legal Track, we analyzed several topics in detail and found that the negotiations often led to a boolean query that was superior to a vector query of the same terms, in various retrieval metrics. However, on average, the final negotiated boolean query matched less than half of the judged relevant items.

In both the Legal and Terabyte Tracks, we looked at both robust and non-robust measures, and they produced the anticipated opposing conclusions about feedback approaches. Robust metrics (such as GenS@10) favor negative feedback techniques, such as duplicate filtering, while non-robust metrics (such as MAP, R-Prec and P10) favor positive feedback techniques, such as pseudo-relevance feedback (blind feedback). For most ad hoc tasks, we would like the organizers to use a robust metric (such as GenS@10) as the main measure. But robust metrics are not appropriate for legal discovery because filtering suspected duplicates is not appropriate in this task.

In our depth probe experiments (depth 9000 in the Legal Track and depth 5000 in the Terabyte Track), we demonstrated that, on average, less than 20% of the relevant items are assessed for the Legal topics and less than 40% of the relevant items are assessed for the Terabyte adhoc topics. Relevant items were found to occur deeply in the results, implying that sampling techniques should look deeper than 9000 rows.

References

[1] 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.

[2] Chris Buckley and Ellen M. Voorhees. Retrieval Evaluation with Incomplete Information. *SIGIR 2004*.

[3] Harr Chen and David R. Karger. Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. *SIGIR 2006*, pp. 429-436.

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

[5] Larry Fitzpatrick, Mei Dent and Gary Promhouse. Experiments with TREC using the Open Text Livelink Engine. Proceedings of TREC-5, 1997.

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

[7] 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.

[8] X. Allan Lu, John D. Holt, David J. Miller. Boolean System Revisited: Its Performance and its Behavior. Proceedings of TREC-4, 1996.

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

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

[11] Stephen Robertson. On GMAP – and other transformations. *CIKM 2006*, pp. 78-83.

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

[13] Stephen Tomlinson. CJK Experiments with Hummingbird SearchServer™ at NTCIR-5. Proceedings of NTCIR-5, 2005.

[14] Stephen Tomlinson. Comparing the Robustness of Expansion Techniques and Retrieval Measures. Working Notes for the CLEF 2006 Workshop.

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

[16] Stephen Tomlinson. Enterprise, QA, Robust and Terabyte Experiments with Hummingbird SearchServer™ at TREC 2005. Proceedings of TREC 2005.

[17] Stephen Tomlinson. European Ad Hoc Retrieval Experiments with Hummingbird SearchServer™ at CLEF 2005. Working Notes for the CLEF 2005 Workshop.

[18] TREC 2006 Terabyte Track Guidelines. http://plg.uwaterloo.ca/∼claclark/TB06.html

[19] TREC Legal Discovery Track: Final Guidelines for 2006 (April 15, 2006). http://trec-legal.umiacs.umd.edu/guidelines6.txt

[20] Howard Turtle. Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance. *SIGIR 1994*, pp. 212-220.

[21] Ellen M. Voorhees. Overview of the TREC 2003 Robust Retrieval Track. Proceedings of TREC 2003.

[22] Ellen M. Voorhees. Overview of the TREC 2004 Robust Retrieval Track. Proceedings of TREC 2004.

[23] Thijs Westerveld, Wessel Kraaij and Djoerd Hiemstra. Retrieving Web Pages using Content, Links, URLs and Anchors. Proceedings of TREC 2001.

[24] Emine Yilmaz and Javed A. Aslam. Estimating Average Precision with Incomplete and Imperfect Judgments. *CIKM 2006*, pp. 102-111.