Enterprise, QA, Robust and Terabyte Experiments with Hummingbird SearchServer$^{\mathrm{TM}}$ at TREC 2005

Stephen Tomlinson
Hummingbird
Ottawa, Ontario, Canada
stephen.tomlinson@hummingbird.com
http://www.hummingbird.com/

February 5, 2006

Abstract

Hummingbird participated in 6 tasks of TREC 2005: the email known-item search task of the Enterprise Track, the document ranking task of the Question Answering Track, the ad hoc topic relevance task of the Robust Retrieval Track, and the adhoc, efficiency and named page finding tasks of the Terabyte Track. In the email known-item task, SearchServer found the desired message in the first 10 rows for more than 80% of the 125 queries. In the document ranking task, SearchServer returned an answering document in the first 10 rows for more than 90% of the 50 questions. In the robustness task, SearchServer found a relevant document in the first 10 rows for 88% of the 50 short (title) topics. In the terabyte adhoc and efficiency tasks, SearchServer found a relevant document in the first 10 rows for more than 90% of the 50 title topics. A new retrieval measure, First Relevant Score, is investigated; it is found to more accurately reflect known-item differences than reciprocal rank and to better reflect robustness across topics than the primary measure of the Robust track.

1 Introduction

Hummingbird SearchServer[1] is a toolkit for developing enterprise search and retrieval applications. The SearchServer kernel is also embedded in other Hummingbird products for the enterprise.

SearchServer works in Unicode internally [4] and supports most of the world's major character sets and languages. The major conferences in text retrieval experimentation (TREC [7], CLEF [3] and NTCIR [5]) 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 enterprise search (known-item search of a specific organization's emails), question answering (finding documents which contain the answer to a question), robust retrieval (robustness of ad hoc search across topics) and terabyte retrieval (adhoc search and named page finding on terabyte scales).

2 Retrieval Measures

Traditionally, different retrieval measures have been used for "ad hoc" tasks, which seek relevant items for a topic, than for "known-item" tasks, which seek a particular known document. However, we argue that

[1] SearchServer$^{\mathrm{TM}}$, SearchSQL$^{\mathrm{TM}}$and Intuitive Searching$^{\mathrm{TM}}$ are trademarks of Hummingbird Ltd. All other copyrights, trademarks and tradenames are the property of their respective owners.

the known-item measures are not only applicable to ad hoc tasks, but that they are often preferable. For many ad hoc tasks, e.g. finding answer documents for questions, just one relevant item is needed. Also, the traditional ad hoc measures encourage retrieval of duplicate relevants, which does not correspond to user benefit.

The traditional known-item measures are very coarse, e.g. Success@10 is 1 or 0 for each topic, while reciprocal rank cannot produce a value between 1.0 and 0.5. This year, we've been investigating a new measure, "First Relevant Score" (defined below), which was introduced in [8]. We consider it our main measure for both ad hoc and known-item tasks.

2.1 Primary Recall Measures

"Primary recall" is retrieval of the first relevant item for a topic. Primary recall measures include the following:

- *First Relevant Score* (FRS): For a topic, FRS is 1.08^{1-r} where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found.
- *Success@n* (S@n): For a topic, Success@n is 1 if a desired page is found in the first n rows, 0 otherwise.
- *Reciprocal Rank* (RR): For a topic, RR is $\frac{1}{r}$ where r is the rank of the first row for which a desired page is found, or zero if a desired page was not found. "Mean Reciprocal Rank" (MRR) is the mean of the reciprocal ranks over all the topics.

Interpretation of FRS: FRS is an estimate of the percentage of potential result list reading the system saved the user to get to the first relevant item, assuming that users are less and less likely to continue reading as they get deeper into the result list.

Comparison of First Relevant Score and Reciprocal Rank: Both FRS and RR are 1.0 if a desired page is found at rank 1. At rank 2, FRS is just 7 points lower (0.93), whereas RR is 50 points lower (0.50). At rank 3, FRS is another 7 points lower (0.86), whereas RR is 17 points lower (0.33). At rank 10, FRS is 0.50, whereas RR is 0.10. FRS is greater than RR for ranks 2 to 52 and lower for ranks 53 and beyond.

Connection of First Relevant Score to Success@10: The base of 1.08 makes FRS 0.5 at rank 10 which in practice makes FRS a good predictor of Success@10 (e.g. if FRS is 0.8, Success@10 will probably be close to 40/50).

This paper lists FRS, S@1, S@10 and MRR for all runs.

2.2 Secondary Recall Measures

"Secondary recall" is retrieval of the additional relevant items for a topic (after the first one). Secondary recall measures place most of their weight on these additional relevant items. They are just considered for ad hoc tasks.

- *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 P20 (Precision@20) for some runs because it was one of the main measures for the Terabyte track.
- *Average Precision* (AP): For a topic, AP is the average of the precision after each relevant document is retrieved (using zero as the precision for relevant documents which are not retrieved). By convention, AP is based on the first 1000 retrieved documents for the topic. 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).

- *Geometric MAP* (GMAP): GMAP was the primary measure for the Robust Track this year (defined in [14]). It is based on "Log Average Precision" which for a topic is the natural log of the sum of 0.00001 and the average precision. GMAP is -0.00001 plus the exponential of the mean log average precision. (We will argue in the Robust section that FRS is a better measure of robustness than the GMAP measure.)
  - The organizers later made a revision to the GMAP definition (which is believed to be minor); this paper just uses the original definition.
- *GMAP'*: We also define a linearized log average precision measure (denoted GMAP') which linearly maps the 'log average precision' values to the [0,1] interval. For statistical significance purposes, GMAP' gives the same results as GMAP, and it has advantages such as that the individual topic differences are in the familiar -1.0 to 1.0 range and are on the same scale as the mean.

2.3 H and J Modifications

We attach an H prefix to the measure (e.g. HFRS, HS@10, HMAP, etc.) when the measure is just counting "highly relevant" documents as relevant. (The H modifier is just applicable to the ad hoc tasks of the Robust and Terabyte tracks, for which the judgements distinguished highly relevants from ordinary relevants.)

We attach a J suffix to the measure (e.g. FRSJ, MAPJ) when unjudged documents are omitted rather than being assumed non-relevant (an approach investigated by [1] for different measures). The J modifier is only applicable for ad hoc tasks, not known-item tasks.

3 Difference Tables

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

- "Expt" specifies the experiment (the codes of the runs being compared are in parentheses).
- "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 experimental run scored higher, lower and tied (respectively) compared to the baseline 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 Enterprise Track: Email Known-Item Search Experiments

The Enterprise Track was new to TREC this year. We participated in its Email Known-Item Search task. The collection to be searched was the 'lists' portion of the "W3C Test Collection". It consisted of the "W3C Public Mailing List Archives" crawled from lists.w3.org in June 2004. Uncompressed, the collection was 1,991,923,793 bytes and consisted of 198,394 documents. The average document size was 10,040 bytes (including HTML markup). For more details, see [2].

For the known-item search queries (e.g. "studies of Web Accessibility for the Disabled"), the goal was to find the particular message the user was trying to retrieve (e.g. "http://lists.w3.org/Archives/Public/w3c-wai-ig/2004AprJun/0111.html"). The organizers provided 25 training queries (for which the right answers were also given) and 125 test queries (for which the right answers were not released until after the submitted runs were due in August 2005). The queries and answers were based on contributions from the participants (including 10 which we contributed). The test queries were numbered from 26 to 150. For more details on the task, see [12].

4.1 Indexing

Our indexing approach was the same as we used in the Web Track each of the previous three years (described in detail in [9]) except that a newer version of the software was used which may have contained an updated English lexicon for stemming.

Briefly: in addition to full-text indexing (except for a short stopword list and some tags), the custom text reader cTREC populated particular columns such as TITLE (if any), URL, META TITLE, META SUBJECT, META DESCRIPTION, META KEYWORDS and the first heading (e.g. first <H1>, <H2> or <H3> content). (Also, URL_TYPE and URL_DEPTH fields were populated, but were not used in the Enterprise experiments.) More details are in [9].

4.2 Searching

The techniques used for the 5 submitted runs of August 2005 (plus one other unsubmitted run produced at that time) are described below. The SearchServer '2:3' relevance method was the same as described last year [11]. Briefly, SearchServer dampens the term frequency and adjusts for document length in a manner similar to Okapi [6] and dampens the inverse document frequency using an approximation of the logarithm. When doing morphological searching (e.g. inflections), these calculations are based on the stems of the terms (roughly speaking).

humEK05l: The submitted humEK05l run was a plain content search including linguistic expansion from English inflectional stemming. This run was the analog of the baseline humR05tl run described in the Robust section (including RELEVANCE_METHOD '2:3' and RELEVANCE_DLEN_IMP 250); the enterprise run used the IS_ABOUT predicate instead of the CONTAINS predicate (and hence the VECTOR_GENERATOR was set to enable inflections instead of the TERM_GENERATOR), but the relevance calculation was the same. This run used almost the same approach as the submitted humW04l run of last year [11] (this year's document length normalization was 250 instead of 500). Below is an example SearchSQL query. Note that the FT_TEXT column indexed the content and also all of the non-content fields of the document source (which included the title and meta tags but not the url):

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM W3CL
WHERE
 (FT_TEXT IS_ABOUT 'studies of Web Accessibility for the Disabled')
ORDER BY REL DESC;
```

humEK05tl: The submitted humEK05tl run was the same as humEK05l except that it put an additional 10% weight on matches in the Title column and an additional 10% weight on phrase matches in the Title (via the CONTAINS predicate). Below is an example SearchSQL query. More details on the syntax are in the description of the humNP03pl run of [10]:

```
WHERE
 (TITLE CONTAINS 'studies of Web Accessibility for the Disabled' WEIGHT 1) OR
 (TITLE IS_ABOUT 'studies of Web Accessibility for the Disabled' WEIGHT 1) OR
 (FT_TEXT IS_ABOUT 'studies of Web Accessibility for the Disabled' WEIGHT 10)
```

Table 1: Mean Scores of Submitted Enterprise Known-Item Search Runs

Run	FRS	S@1	S@5	S@10	S@100	MRR
humEK05t3l	0.774	61/125	96/125	101/125	114/125	0.604
humEK05tl	0.770	58/125	95/125	102/125	115/125	0.590
humEK05pl	0.770	59/125	94/125	101/125	115/125	0.595
(humEK05v3l)	0.769	58/125	95/125	101/125	114/125	0.587
humEK05p	0.730	54/125	88/125	94/125	115/125	0.548
humEK05l	0.718	47/125	87/125	95/125	114/125	0.513

humEK05t3l: The submitted humEK05t3l run was the same as humEK05tl except that it put 3 times more weight on the title fields. Note that the content still had more weight overall. Below is an example SearchSQL query:

```
WHERE
 (TITLE CONTAINS 'studies of Web Accessibility for the Disabled' WEIGHT 3) OR
 (TITLE IS_ABOUT 'studies of Web Accessibility for the Disabled' WEIGHT 3) OR
 (FT_TEXT IS_ABOUT 'studies of Web Accessibility for the Disabled' WEIGHT 10)
```

humEK05pl: The submitted humEK05pl run was the same as humEK05tl except that in place of the TITLE field it searched the ALL_PROPS field which was a combination of the title, url, some meta tags and the first heading. This run was the analog of last year's humW04pl run [11] (except for the lower document length setting this year):

```
WHERE
 (ALL_PROPS CONTAINS 'studies of Web Accessibility for the Disabled' WEIGHT 1) OR
 (ALL_PROPS IS_ABOUT 'studies of Web Accessibility for the Disabled' WEIGHT 1) OR
 (FT_TEXT IS_ABOUT 'studies of Web Accessibility for the Disabled' WEIGHT 10)
```

humEK05p: The submitted humEK05p run was the same as humEK05pl except that inflections from stemming were disabled with SET VECTOR_GENERATOR '' (which disables inflections for the IS_ABOUT predicate). Note that inflections were not enabled for the CONTAINS predicate for any Enterprise run.

humEK05v3l: (This run was not submitted.) The humEK05v3l run was the same as humEK05t3l except that the extra weight on a phrase match in the title was removed:

```
WHERE
 (TITLE IS_ABOUT 'studies of Web Accessibility for the Disabled' WEIGHT 3) OR
 (FT_TEXT IS_ABOUT 'studies of Web Accessibility for the Disabled' WEIGHT 10)
```

4.3 Results

Table 1 lists the mean scores of the 5 submitted runs (plus 1 extra run). The baseline full-text search technique returned the desired message in the first 10 rows for 76% of the 125 test queries (95/125). With additional weight on the Title field, Success@10 increased to 82% (102/125); this increase was statistically significant according to Table 2 (i.e. the approximate 95% confidence interval of the "t (tl-l)" line of the ΔS10 section does not contain zero).

- Title weighting: Table 2 shows that the increase in First Relevant Score from the extra 10% weights on the title ("t" experiment) was sometimes dramatic, e.g. topic KI107 increased in score from 0.00 to 1.00. The negative impacts for FRS in the "t" experiment were not large (the biggest was a 21-point

Table 2: Impact of Enterprise Known-Item Search Techniques

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
t3 (t3l-l)	0.056	(0.018, 0.094)	39-18-68	1.00 (107), 0.96 (117), −0.69 (131)
t (tl-l)	0.052	(0.019, 0.086)	33-10-82	1.00 (107), 0.96 (117), −0.21 (131)
p (pl-l)	0.052	(0.018, 0.087)	35-10-80	1.00 (107), 0.96 (117), −0.45 (131)
v3 (v3l-l)	0.051	(0.015, 0.087)	36-18-71	1.00 (107), 0.96 (117), −0.68 (131)
l (pl-p)	0.040	(0.003, 0.078)	21-15-89	1.00 (79), 1.00 (124), −0.52 (131)
	ΔS1			
t3 (t3l-l)	0.112	(0.035, 0.189)	19-5-101	1.00 (67), 1.00 (108), −1.00 (56)
p (pl-l)	0.096	(0.030, 0.162)	15-3-107	1.00 (79), 1.00 (108), −1.00 (47)
t (tl-l)	0.088	(0.023, 0.153)	14-3-108	1.00 (79), 1.00 (67), −1.00 (47)
v3 (v3l-l)	0.088	(0.023, 0.153)	14-3-108	1.00 (52), 1.00 (74), −1.00 (47)
l (pl-p)	0.040	(−0.022, 0.102)	10-5-110	1.00 (79), 1.00 (112), −1.00 (150)
	ΔS10			
t (tl-l)	0.056	(0.014, 0.098)	7-0-118	1.00 (71), 1.00 (108), 0.00 (150)
p (pl-l)	0.048	(0.003, 0.093)	7-1-117	1.00 (71), 1.00 (108), −1.00 (131)
l (pl-p)	0.056	(−0.002, 0.114)	10-3-112	1.00 (71), 1.00 (146), −1.00 (116)
t3 (t3l-l)	0.048	(−0.003, 0.099)	8-2-115	1.00 (71), 1.00 (108), −1.00 (66)
v3 (v3l-l)	0.048	(−0.003, 0.099)	8-2-115	1.00 (71), 1.00 (108), −1.00 (66)
	ΔMRR			
t3 (t3l-l)	0.091	(0.036, 0.146)	39-18-68	1.00 (107), 0.98 (117), −0.89 (148)
p (pl-l)	0.082	(0.032, 0.131)	35-10-80	1.00 (107), 0.98 (117), −0.80 (148)
t (tl-l)	0.076	(0.028, 0.125)	33-10-82	1.00 (107), 0.98 (117), −0.75 (148)
v3 (v3l-l)	0.074	(0.026, 0.121)	36-18-71	1.00 (107), 0.98 (117), −0.86 (148)
l (pl-p)	0.047	(−0.004, 0.097)	21-15-89	1.00 (79), 0.99 (124), −0.83 (67)

decrease on topic KI131 according to Table 2, a fall from rank 4 to 8). The other title-weighting techniques ("p", "v3" and "t3") had larger negative per-topic impacts, but overall were pretty similar. It appears a small extra weight on the title is a reasonable general-purpose technique.

- Inflections from stemming: Table 2 shows that the increase in First Relevant Score from inflections from stemming ("l" experiment) was also sometimes dramatic, though there could also be large decreases on some topics. The mean increase in FRS was statistically significant (the increase did not quite pass the significance test for the other measures in Table 2). Stemming may be a reasonable default behaviour, but the interface should probably give a user a way to disable it for particular terms.

One can see the advantages of First Relevant Score compared to Reciprocal Rank in Table 2. For example, for the "l" experiment, reciprocal rank picks topic 67 as the largest decrease (83 points), though it is just a fall from rank 1 to 6; FRS considers this just a 32 point decrease. FRS picks topic 131 as the largest decrease (52 points), a bigger fall from rank 3 to 15; RR considers this just a 27-point decrease. The confidence intervals for FRS are also narrower than for MRR.

One can see in Table 1 that (mean) FRS was a good predictor of (mean) Success@10. Also, in Table 2, the confidence intervals for the mean change in FRS are fairly similar to (but also narrower than) those for S@10.

5 Question Answering Track: Document Ranking Experiments

In the Document Ranking task of the Question Answering Track, the collection to be searched was the "AQUAINT collection" of English newswire articles from the 1998-2000 time period. Uncompressed, the collection was 3,181,313,864 bytes and consisted of 1,033,461 documents. The average document size was 3078 bytes (including SGML markup).

Each test question included a “target” (e.g. “skier Alberto Tomba”) and the question itself (e.g. “What nationality is he?”). The goal was to find all the documents which contained the answer to the question and return them at the top of the list.

The organizers provided 50 test questions. However, some of them were from the same series (from another Question Answering task) and hence shared the same target. So unlike for most TREC tasks, the 50 queries likely were not independent, making the mean scores less reliable and invalidating the usual approach to statistical significance testing.

To create a test set with independent questions, we just kept the first judged question for each target, discarding additional questions from the same series. This revised test set contained 38 questions. Our diagnostics are on this latter set.

Over the original 50 test questions, there were on average 31.5 answer documents per question (low 1, high 285, median 6.5). Over the independent 38 test question subset, there were on average 34 answer documents per question (low 1, high 285, median 7.5).

5.1 Indexing

The indexing approach was mostly the same as for the Robust Retrieval Track of last year [11]. We used a SearchServer index which supported both exact matching (after some Unicode-based normalizations, such as decompositions and conversion to upper-case) and matching of inflections based on English lexical stemming (i.e. stemming based on a dictionary or lexicon for the language). For example, in English, “baby”, “babied”, “babies”, “baby’s” and “babying” all have “baby” as a stem. Some stop words were excluded from indexing (e.g. “the”, “by” and “of”). Based on looking at the questions from the previous year’s QA Track, we added a few more stopwords (‘how’, ‘many’ and ‘kind’).

5.2 Searching

The techniques used for the 3 submitted runs of July 2005 are described below. The base technique was to use the SearchServer CONTAINS predicate to perform a boolean-OR of the words of the target and question.

humQ05l: The submitted humQ05l run was a plain content search including linguistic expansion from English inflectional stemming. This run was the same approach as last year’s humR04d5 run (including RELEVANCE_METHOD ‘2:3’ and RELEVANCE_DLEN_IMP 500) except that instruction words (such as “find”, “relevant” and “document”) were not removed. Below is an example SearchSQL query:

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM AQ05
WHERE
 FT_TEXT CONTAINS 'skier'|'Alberto'|'Tomba'|'What'|'nationality'|'is'|'he'
ORDER BY REL DESC;
```

humQ05xl: The submitted humQ05xl run was the same as humQ05l 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). Below is an example WHERE clause of a SearchSQL query:

```
WHERE
 FT_TEXT CONTAINS 'skier' WEIGHT 5|'Alberto' WEIGHT 5|'Tomba' WEIGHT 5|
                  'What' WEIGHT 5|'nationality' WEIGHT 5|
                  'is' WEIGHT 5|'he' WEIGHT 5
 OR FT_TEXT CONTAINS PROXIMITY 200 CHARACTERS
  ('skier'&'Alberto'&'Tomba'&'What'&'nationality'&'is'&'he')
```

Table 3: Mean Scores of Submitted Question Answering Runs (50 and 38 Question Sets)

Run	FRS	S@1	S@10	P20	MRR	MAP	FRSJ	MAPJ
humQ05xl	0.882	33/50	46/50	0.301	0.749	0.416	0.882	0.426
humQ05l	0.880	34/50	46/50	0.286	0.753	0.413	0.881	0.426
humQ05xle	0.883	33/50	45/50	0.315	0.750	0.447	0.884	0.459
[independent 38]	FRS	S@1	S@10	P20	MRR	MAP	FRSJ	MAPJ
humQ05xl	0.866	23/38	35/38	0.275	0.715	0.389	0.866	0.398
humQ05l	0.864	25/38	35/38	0.267	0.732	0.391	0.864	0.400
humQ05xle	0.857	23/38	33/38	0.287	0.710	0.420	0.858	0.431

Table 4: Impact of Question Answering Techniques (38 Question Set)

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Series)
x (xl-l)	0.001	(−0.009, 0.011)	3-2-33	0.13 (73), −0.07 (101), −0.07 (127)
e (xle-xl)	−0.009	(−0.041, 0.024)	4-8-26	0.39 (96), −0.27 (107), −0.33 (136)
	ΔMRR			
x (xl-l)	−0.018	(−0.058, 0.023)	3-2-33	−0.50 (127), −0.50 (101), 0.25 (73)
e (xle-xl)	−0.005	(−0.060, 0.050)	4-8-26	−0.50 (123), −0.50 (89), 0.50 (73)
	ΔMAP			
x (xl-l)	−0.002	(−0.020, 0.017)	14-17-7	−0.25 (127), −0.10 (78), 0.09 (135)
e (xle-xl)	0.031	(−0.002, 0.063)	21-14-3	0.38 (73), 0.26 (113), −0.15 (104)

humQ05xle: The submitted humQ05xle run was a blind feedback run based 50% on humQ05xl and 25% each on expansion queries from the first 2 rows of humQ05xl. (The expansion queries used a document length normalization of 750.) If the base run contains answer documents at the top of the list, this approach is likely to find more answer documents. From a user perspective, padding the results with more answer documents is unimportant because one answer document presumably should suffice, but the organizers wanted to focus on the recall-oriented 'mean average precision' measure which usually benefits from blind feedback approaches.

5.3 Results

Table 3 lists the mean scores of the 3 submitted runs. Each run returned an answer document in the first 10 rows for at least 90% of the questions (except for the blind feedback run on the 38-question subset). The proximity technique made little difference on average. The mean differences for blind feedback also did not quite pass the significance test on the 38-question subset (as per Table 4).

Table 5 lists the mean scores of the diagnostic runs. For each approach (boolean-OR and boolean-AND), there was a "baseline" run (SET RELEVANCE_METHOD '2:3', SET RELEVANCE_DLEN_IMP 250, SET TERM_GENERATOR 'word!ftelp/inflect'). These settings were the same as for the humQ05l run except that RELEVANCE_DLEN_IMP was set to 250 instead of 500. (No proximity nor blind feedback was used for these diagnostic runs.) The other runs just had the one listed difference from the baseline run (these differences are explained in detail in Section 2.3 of last year's paper [11]).

Table 6 isolates the differences from the baseline runs. Document length normalization and stemming both had statistically significant mean benefits for the first relevant item. Switching to the '2:4' method (squaring the importance of inverse document frequency (idf)) did not make a statistically significant difference, but disabling idf completely ('2:5') appeared to be detrimental. Like last year, the hits count method ('2:1') was not competitive for boolean-OR, but was respectable for boolean-AND.

Table 5: Mean Scores of Diagnostic Question Answering Runs (38 Question Set)

Run	FRS	S@1	S@10	P20	MRR	MAP	FRSJ	MAPJ
OR: “2:3” (normal idf)	0.872	27/38	34/38	0.262	0.770	0.401	0.872	0.412
OR: “2:3 with no dlen”	0.838	22/38	34/38	0.239	0.684	0.344	0.838	0.359
OR: “2:4” (idf squared)	0.836	25/38	33/38	0.257	0.731	0.379	0.836	0.390
OR: “2:3 with no stemming”	0.835	22/38	34/38	0.249	0.677	0.345	0.836	0.357
OR: “2:5” (no idf)	0.813	21/38	33/38	0.234	0.653	0.308	0.821	0.329
OR: “2:2” (terms count)	0.488	8/38	17/38	0.147	0.306	0.137	0.523	0.166
OR: “2:1” (hits count)	0.376	9/38	14/38	0.079	0.291	0.100	0.516	0.145
AND: “2:3” (normal idf)	0.701	24/38	27/38	0.172	0.664	0.213	0.701	0.214
AND: “2:4” (idf squared)	0.697	23/38	27/38	0.175	0.647	0.200	0.697	0.202
AND: “2:5” (no idf)	0.670	19/38	27/38	0.171	0.568	0.183	0.670	0.186
AND: “2:3 with no dlen”	0.659	21/38	26/38	0.168	0.586	0.180	0.664	0.183
AND: “2:1” (hits count)	0.626	19/38	25/38	0.149	0.536	0.159	0.634	0.165
AND: “2:3 with no stemming”	0.548	16/38	22/38	0.117	0.475	0.123	0.548	0.124
AND: “2:2” (terms count)	0.445	8/38	15/38	0.134	0.295	0.094	0.467	0.101

Table 6: Impact of Diagnostic Question Answering Techniques (38 Question Set)

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Series)
OR: no dlen	−0.034	(−0.066,−0.002)	2-10-26	−0.37 (73), −0.26 (120), 0.19 (136)
OR: 2:4	−0.036	(−0.092, 0.020)	4-6-28	−0.95 (89), −0.40 (73), 0.13 (91)
OR: no stem	−0.037	(−0.074, 0.000)	4-10-24	−0.50 (89), −0.32 (125), 0.15 (96)
OR: 2:5	−0.059	(−0.120, 0.002)	2-13-23	−0.79 (107), −0.63 (136), 0.29 (108)
OR: 2:2	−0.385	(−0.498,−0.271)	0-29-9	−1.00 (135), −1.00 (89), 0.00 (78)
OR: 2:1	−0.496	(−0.641,−0.352)	1-28-9	−1.00 (120), −1.00 (127), 0.07 (126)
AND: 2:4	−0.004	(−0.014, 0.006)	1-2-35	−0.14 (73), −0.07 (125), 0.07 (101)
AND: 2:5	−0.031	(−0.056,−0.006)	0-7-31	−0.30 (104), −0.26 (79), 0.00 (108)
AND: no dlen	−0.042	(−0.082,−0.002)	0-5-33	−0.53 (104), −0.42 (73), 0.00 (108)
AND: 2:1	−0.076	(−0.133,−0.018)	1-8-29	−0.66 (104), −0.54 (73), 0.07 (101)
AND: no stem	−0.154	(−0.267,−0.040)	1-9-28	−1.00 (131), −1.00 (66), 0.13 (109)
AND: 2:2	−0.256	(−0.367,−0.145)	0-19-19	−1.00 (135), −0.95 (92), 0.00 (108)
	ΔMAP			
OR: 2:4	−0.021	(−0.066, 0.023)	15-20-3	−0.67 (125), −0.34 (89), 0.26 (97)
OR: no stem	−0.056	(−0.109,−0.003)	15-20-3	−0.83 (125), −0.35 (84), 0.15 (109)
OR: no dlen	−0.057	(−0.102,−0.012)	7-28-3	−0.80 (120), −0.23 (89), 0.05 (136)
OR: 2:5	−0.092	(−0.137,−0.048)	6-30-2	−0.67 (120), −0.38 (75), 0.08 (108)
OR: 2:2	−0.264	(−0.344,−0.184)	1-36-1	−0.93 (74), −0.92 (125), 0.01 (105)
OR: 2:1	−0.301	(−0.382,−0.220)	1-35-2	−1.00 (120), −1.00 (74), 0.08 (126)
AND: 2:4	−0.012	(−0.040, 0.015)	9-8-21	−0.50 (125), −0.11 (73), 0.06 (109)
AND: 2:5	−0.029	(−0.067, 0.008)	5-13-20	−0.67 (120), −0.16 (98), 0.05 (93)
AND: no dlen	−0.032	(−0.075, 0.011)	4-15-19	−0.80 (120), −0.15 (73), 0.03 (84)
AND: 2:1	−0.054	(−0.105,−0.002)	4-16-18	−0.86 (120), −0.42 (127), 0.05 (93)
AND: no stem	−0.090	(−0.159,−0.021)	2-20-16	−1.00 (120), −0.80 (125), 0.15 (109)
AND: 2:2	−0.119	(−0.192,−0.047)	3-20-15	−0.92 (125), −0.86 (120), 0.05 (79)

Overall, this year's diagnostic results for finding answer documents were pretty similar to last year's diagnostic results for finding relevant documents.

6 Robust Retrieval Track: Topic Relevance Experiments

The Robust Retrieval Track used the same AQUAINT collection described in the Question Answering section. But instead of finding answer documents for questions, the goal was to find relevant documents for topics.

Each topic contained a "Title" (subject of the topic, e.g. "killer bee attacks"), "Description" (a one-sentence specification of the information need, e.g. "Identify instances of attacks on humans by Africanized (killer) bees.") and "Narrative" (more detailed guidelines for what a relevant document should or should not contain, e.g. "Relevant documents must cite a specific instance of a human attacked by killer bees. Documents that note migration patterns or report attacks on other animals are not relevant unless they also cite an attack on a human.").

The organizers actually re-used 50 of the more "difficult" topics from past TREC ad hoc tracks, but the document set was different, so new relevance assessments were done (and some concern was expressed that the judging pools were shallower than past years).

The judgements contained on average 131 relevant documents per topic (low 9, high 376, median 113) counting both "relevant" and "highly relevant" as relevant. If just "highly relevants" are counted as relevant, then 5 topics are discarded (for having no highly relevants), and over the remaining 45 topics, there were 62 highly relevants per topic (low 1, high 334, median 50).

6.1 Indexing

The same index was used as described in the Question Answering section.

6.2 Searching

The techniques used for the 5 submitted runs of July 2005 (and 2 other runs produced at the same time) are described below. The base technique was to use the SearchServer CONTAINS predicate to perform a boolean-OR of the words of the topic field.

humR05tl: The submitted humR05tl run was a plain content search (including linguistic expansion from English inflectional stemming) on the Title field of the topic. This run used the same approach as the humQ05l run except that we used RELEVANCE_DLEN_IMP 250 for the title runs. Below is an example SearchSQL query:

```
SELECT RELEVANCE('2:3') AS REL, DOCNO
FROM AQ05
WHERE
 FT_TEXT CONTAINS 'killer'|'bee'|'attacks'
ORDER BY REL DESC;
```

humR05txl: The submitted humR05txl run was the same as humR05tl 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). This run used the same approach as the humQ05xl run except that we used RELEVANCE_DLEN_IMP 250 for the title runs. Below is an example WHERE clause of a SearchSQL query:

```
WHERE
 FT_TEXT CONTAINS 'killer' WEIGHT 5|'bee' WEIGHT 5|'attacks' WEIGHT 5
 OR FT_TEXT CONTAINS PROXIMITY 200 CHARACTERS ('killer'&'bee'&'attacks')
```

Table 7: Mean Scores of Submitted Robust Retrieval Runs

Run	FRS	S@1	S@10	MRR	MAP	GMAP	FRSJ	MAPJ
humR05tl	0.833	27/50	44/50	0.653	0.202	0.126	0.833	0.237
humR05txl	0.831	28/50	44/50	0.663	0.208	0.132	0.835	0.245
(humR05tx5l)	0.824	24/50	44/50	0.608	0.195	0.122	0.830	0.238
humR05txle	0.795	26/50	42/50	0.629	0.242	0.150	0.797	0.272
humR05dl	0.801	22/50	42/50	0.577	0.158	0.091	0.817	0.212
(humR05dxl)	0.788	20/50	42/50	0.546	0.158	0.092	0.806	0.213
humR05dle	0.752	22/50	38/50	0.561	0.201	0.114	0.771	0.244
[on highly rels]	HFRS	HS@1	HS@10	HMRR	HMAP	HGMAP	HFRSJ	HMAPJ
humR05tl	0.659	13/45	32/45	0.420	0.123	0.050	0.659	0.139
humR05txl	0.651	11/45	32/45	0.398	0.124	0.051	0.654	0.141
(humR05tx5l)	0.626	11/45	33/45	0.371	0.110	0.045	0.631	0.129
humR05txle	0.613	11/45	30/45	0.380	0.141	0.054	0.616	0.154
humR05dl	0.603	10/45	28/45	0.352	0.096	0.043	0.619	0.123
(humR05dxl)	0.598	9/45	28/45	0.338	0.096	0.043	0.614	0.124
humR05dle	0.548	10/45	24/45	0.329	0.125	0.048	0.563	0.145

humR05tx5l: (This run was not submitted.) The humR05tx5l run was the same as humR05txl except 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.) Below is an example WHERE clause of a SearchSQL query:

```
WHERE
 FT_TEXT CONTAINS ’killer’|’bee’|’attacks’
 OR FT_TEXT CONTAINS PROXIMITY 200 CHARACTERS
    (’killer’ WEIGHT 5 & ’bee’ WEIGHT 5 & ’attacks’ WEIGHT 5)
```

humR05txle: The submitted humR05txle run was a blind feedback run based 50% on humR05txl and 25% each on expansion queries from the first 2 rows of humR05txl. (The expansion queries used a document length normalization of 750.)

humR05dl: The submitted humR05dl run used the same approach as humR05tl except that the Description field of the topic was used instead of the Title, instruction words such as “find”, “relevant” and “document” were discarded before forming the query (same list as last year), and RELEVANCE_DLEN_IMP 500 was used for the description runs. This run used the same approach as the humR04d5 run of last year.

humR05dxl: (This run was not submitted.) The humR05dxl run was to the humR05dl run as humR05txl was to humR05tl.

humR05dle: The submitted humR05dle run was a blind feedback run based 50% on humR05dl and 25% each on expansion queries from the first 2 rows of humR05dl. (The expansion queries used a document length normalization of 750.)

For the submitted runs, the participants were asked to append a ranking of the system’s confidence of how well it did on the topic. For each run, we just used the relevance value (i.e. the number returned by the SearchServer RELEVANCE() function) of the top-retrieved row as our basis for ranking the topics. A higher relevance value was considered to mean a higher confidence in the relevance of the document. (This was the same approach as we used last year, and Section 2.4.1 of last year’s paper [11] analyzed the results.)

Table 8: Impact of Robust Retrieval Techniques

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
x (txl-tl)	−0.001	(−0.034, 0.031)	7-8-35	0.59 (322), −0.20 (448), −0.39 (383)
x5 (tx5l-tl)	−0.008	(−0.042, 0.025)	10-15-25	0.46 (426), 0.41 (322), −0.26 (439)
X (dxl-dl)	−0.012	(−0.024,−0.001)	1-7-42	−0.19 (416), −0.18 (419), 0.01 (448)
e (txle-txl)	−0.037	(−0.074, 0.001)	5-12-33	−0.83 (322), −0.26 (367), 0.10 (378)
E (dle-dl)	−0.049	(−0.096,−0.002)	7-14-29	−0.42 (341), −0.42 (435), 0.40 (448)
	ΔHFRS			
X (dxl-dl)	−0.005	(−0.016, 0.006)	5-7-33	−0.18 (419), −0.07 (426), 0.10 (650)
x (txl-tl)	−0.008	(−0.054, 0.038)	8-16-21	0.59 (322), −0.32 (689), −0.39 (383)
x5 (tx5l-tl)	−0.033	(−0.093, 0.026)	8-22-15	−0.46 (409), −0.46 (658), 0.46 (426)
e (txle-txl)	−0.038	(−0.083, 0.008)	8-16-21	−0.83 (322), −0.31 (344), 0.22 (307)
E (dle-dl)	−0.055	(−0.119, 0.009)	11-17-17	−0.52 (372), −0.46 (439), 0.48 (416)
	ΔMAP			
E (dle-dl)	0.044	(0.025, 0.062)	38-12-0	0.23 (622), 0.20 (625), −0.10 (408)
e (txle-txl)	0.034	(0.015, 0.052)	35-15-0	0.22 (325), 0.20 (622), −0.11 (303)
x (txl-tl)	0.007	(−0.002, 0.016)	26-22-2	0.18 (648), 0.06 (394), −0.04 (325)
X (dxl-dl)	−0.000	(−0.002, 0.002)	23-23-4	0.03 (393), −0.01 (416), −0.02 (394)
x5 (tx5l-tl)	−0.006	(−0.022, 0.009)	21-28-1	0.21 (648), 0.08 (336), −0.21 (374)
	ΔHMAP			
E (dle-dl)	0.029	(0.009, 0.049)	27-17-1	0.24 (622), 0.20 (374), −0.10 (372)
e (txle-txl)	0.017	(0.000, 0.033)	21-21-3	0.25 (622), 0.15 (625), −0.08 (303)
x (txl-tl)	0.001	(−0.007, 0.010)	18-23-4	0.14 (648), −0.05 (374), −0.06 (650)
X (dxl-dl)	0.001	(−0.001, 0.003)	23-15-7	0.03 (393), 0.01 (650), −0.01 (419)
x5 (tx5l-tl)	−0.012	(−0.031, 0.006)	13-29-3	−0.30 (374), −0.09 (650), 0.17 (648)
	ΔGMAP’			
E (dle-dl)	0.019	(0.007, 0.031)	38-12-0	0.13 (336), 0.12 (345), −0.12 (651)
e (txle-txl)	0.011	(0.002, 0.021)	35-15-0	−0.09 (651), 0.08 (372), 0.08 (625)
x (txl-tl)	0.004	(0.000, 0.008)	26-22-2	0.04 (651), 0.04 (648), −0.02 (448)
X (dxl-dl)	0.000	(−0.002, 0.002)	23-23-4	0.03 (393), −0.01 (419), −0.01 (426)
x5 (tx5l-tl)	−0.002	(−0.009, 0.004)	21-28-1	0.06 (393), 0.04 (648), −0.06 (650)
	ΔHGMAP’			
E (dle-dl)	0.009	(−0.008, 0.026)	27-17-1	−0.17 (651), −0.12 (372), 0.15 (336)
e (txle-txl)	0.005	(−0.009, 0.018)	21-21-3	−0.15 (322), −0.10 (651), 0.10 (625)
x (txl-tl)	0.003	(−0.005, 0.011)	18-23-4	0.11 (322), 0.06 (330), −0.05 (650)
X (dxl-dl)	0.001	(−0.001, 0.003)	23-15-7	0.03 (393), −0.01 (419), −0.01 (426)
x5 (tx5l-tl)	−0.008	(−0.019, 0.003)	13-29-3	−0.11 (650), −0.07 (374), 0.07 (330)

6.3 Results

Even though the topics were chosen in part because they had been difficult in past tracks, Table 7 shows that on short title queries, the plain content search technique returned a relevant document in the first 10 rows for 88% of the topics (44/50). When available, a highly relevant was returned in the first 10 rows for 71% of the topics (32/45).

- ‘x’ experiment: The modest proximity weighting technique for titles (‘x’ experiment in Table 8) led to a borderline significant increase in GMAP, though the individual impacts on average precision were small. The biggest increase for GMAP was the increase in average precision from 0.02 to 0.03 for topic 651 (“U.S. ethnic population”), followed by the increase in average precision from 0.32 to 0.50 for topic

648 ("family leave law"); the latter was the largest increase for the MAP measure. Proximity had bigger impacts on First Relevant Score, but was neutral on average.

- 'X' experiment: On descriptions, the modest proximity weighting ('X' experiment in Table 8) actually led to a borderline significant decrease in FRS. For example, for topic 416 ("What is the status of The Three Gorges Project?"), the first relevant fell from rank 2 to rank 5, apparently because the proximity technique required all non-stop words to be close together to get extra weight, and the word "status" wasn't helpful in this case. For some other descriptions, no documents matched the proximity clause, but the integer relevance scores from the OR-matches were squeezed into a smaller range, and the extra ties apparently caused a minor degrading of the ranking.

- 'x5' experiment: The heavier weight on proximity weighting for titles ('x5' experiment in Table 8) tended to be detrimental on average (though the mean differences were not statistically significant). There were some large per-topic impacts in each direction.

- 'e' experiment: The blind feedback technique for titles ('e' experiment in Table 8) produced a statistically significant increase in MAP and GMAP, but FRS was negatively impacted.

- 'E' experiment: On descriptions, the blind feedback technique ('E' experiment in Table 8) produced statistically significant changes in *opposite* directions: increases for MAP and GMAP, but a decrease for FRS.

One can see why we don't hear of the "blind feedback" technique being used in practice. When searching for one item, it is detrimental. When a lot of relevant documents will be reviewed, it's worth your time to supervise the query enhancement process. Either way, the "blind" form of feedback does not make sense in practice.

Table 8 shows that the mean differences for the track's new GMAP measure correlated strongly with those for MAP, even though GMAP picked out different extreme per-topic differences than MAP. In particular, GMAP still significantly favored the non-robust blind feedback technique.

GMAP was supposed to emphasize poorly performing topics, so why did it fail? On topic 341, for which blind feedback ('E' experiment) caused FRS to fall 42 points (the first relevant fell from rank 5 to 16), average precision actually increased slightly (from 0.0247 to 0.0252), so GMAP also increased slightly. On topic 435, for which blind feedback caused FRS to fall 42 points (again the first relevant fell from rank 5 to 16), average precision fell slightly from 0.0320 to 0.0263, and while GMAP' fell by more (from 0.70 to 0.68 in the linearized version), it still wasn't a substantial drop. A topic GMAP' did emphasize a lot was topic 345, for which average precision increased from 0.0018 to 0.0070, which GMAP' considered a substantial improvement (from 0.45 to 0.57 in the linearized scores), but the first relevant moved up from just (approximately) rank 345 to rank 127, and it doesn't seem that a user would consider this so substantial (FRS considers it just a 0.0001 increase).

In our poster at the TREC conference, we looked at using blind feedback as a "litmus test" for robustness measures:

- Blind feedback produced statistically significant increases for MAP, GMAP, R-Precision, Precision@20 and Interpolated Precision at 10% Recall, suggesting that these measures are not suitable as robustness measures.

- Blind feedback produced statistically significant decreases for FRS, MRR and Success@10, suggesting that these measures are candidates as robustness measures.

It appears that "primary recall" measures reflect robustness, while "secondary recall" measures do not.

Putting aside the notion of robustness, it's clear that primary and secondary recall measures can have opposite conclusions about a technique. Unfortunately, most past studies of ad hoc search have only reported secondary measures. Potentially a lot of "established" results for ad hoc search do not apply to retrieval of the first relevant item, particularly those involving blind feedback techniques.

7 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 [13].

7.1 Indexing

The indexing approach was the same as described in the Enterprise section.

In the terabyte adhoc and efficiency tasks, the searches did not make use of the extra columns (e.g. title, meta tags, etc.). The full-text of the document (FT_TEXT column), however, in the case of html documents included the title, strings from the meta tags, etc., though not the url.

In the terabyte named page finding searches, some of the submitted runs used not only the same extra columns as some of the Enterprise runs (title, meta tags, etc.), but also the url type and depth columns. The URL_TYPE was set to ROOT, SUBROOT, PATH or FILE, based on the convention which worked well in TREC 2001 for the Twente/TNO group [15] on the entry page finding task (also known as the home page finding task). The URL_DEPTH was set to a term indicating the depth of the page in the site. Examples and the exact rules we used are given in [9].

Unlike for last year's submitted terabyte runs, the entire collection was indexed in one SearchServer table.

7.2 Adhoc Experiments

The Adhoc Task of the Terabyte Track was much like the topic relevance task of the Robust Retrieval Track. As in the Robust task, there were 50 topics, each with a title, description and narrative field. In the Terabyte adhoc task, the topics were new and meant to be typical (rather than re-using old topics which had been difficult in the past). And of course, the Terabyte task was searching a collection with 24 times as many documents as the Robust task (and the documents were more than 5 times longer on average and were from government web sites rather than news articles).

The terabyte adhoc judgements contained on average 208 relevant documents per topic (low 4, high 559, median 172) counting both "relevant" and "highly relevant" as relevant. If just "highly relevants" are counted as relevant, then 3 topics are discarded (for having no highly relevants), and over the remaining 47 topics, there were 56 highly relevants per topic (low 1, high 331, median 36).

The techniques used for the 4 submitted runs of July 2005 are described below. The base technique was to use the SearchServer CONTAINS predicate to perform a boolean-OR of the words of the Title field of the topic.

humT05l: The submitted humT05l was a plain content search (including linguistic expansion from English inflectional stemming) on the Title field of the topic. This run used the same approach as the humR05tl run described in the Robust section.

humT05xl: The submitted humT05xl run was the same as humT05l 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). This run used the same approach as the humR05txl run described in the Robust section (which has example query syntax).

humT05x5l: The submitted humR05tx5l run was the same as humT05xl 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.) This run used the same approach as the humR05tx5l run described in the Robust section (which has example query syntax).

Table 9: Mean Scores of Submitted Terabyte Adhoc Runs

Run	FRS	S@1	S@10	P20	MRR	MAP	FRSJ	MAPJ
humT05x5l	0.932	36/50	48/50	0.565	0.815	0.332	0.932	0.376
humT05l	0.930	35/50	47/50	0.580	0.807	0.315	0.930	0.356
humT05xl	0.929	38/50	48/50	0.596	0.826	0.336	0.929	0.374
humT05xle	0.903	35/50	46/50	0.623	0.780	0.365	0.904	0.413
(on highly rels)	HFRS	HS@1	HS@10	HP20	HMRR	HMAP	HFRSJ	HMAPJ
humT05xl	0.711	20/47	35/47	0.232	0.526	0.208	0.711	0.218
humT05l	0.694	19/47	33/47	0.222	0.519	0.195	0.694	0.206
humT05x5l	0.689	18/47	35/47	0.206	0.492	0.207	0.689	0.218
humT05xle	0.660	16/47	32/47	0.232	0.470	0.222	0.661	0.236

Table 10: Impact of Terabyte Adhoc Techniques

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
x5 (x5l-l)	0.002	(−0.033, 0.037)	6-7-37	0.56 (762), 0.32 (792), −0.39 (763)
x (xl-l)	−0.001	(−0.023, 0.022)	6-5-39	−0.30 (763), −0.25 (777), 0.25 (792)
e (xle-xl)	−0.026	(−0.056, 0.003)	4-11-35	−0.46 (754), −0.46 (769), 0.20 (795)
	ΔHFRS			
x (xl-l)	0.017	(−0.011, 0.044)	12-9-26	0.39 (783), 0.25 (798), −0.30 (763)
x5 (x5l-l)	−0.005	(−0.072, 0.061)	8-17-22	0.63 (760), 0.57 (783), −0.58 (777)
e (xle-xl)	−0.051	(−0.107, 0.005)	11-18-18	−0.79 (775), −0.63 (792), 0.27 (777)
	ΔMAP			
e (xle-xl)	0.029	(0.004, 0.055)	30-20-0	0.39 (773), 0.25 (772), −0.14 (770)
x (xl-l)	0.021	(0.012, 0.029)	35-13-2	0.11 (785), 0.08 (772), −0.02 (756)
x5 (x5l-l)	0.017	(−0.016, 0.050)	24-25-1	0.54 (785), 0.21 (755), −0.18 (777)
	ΔHMAP			
e (xle-xl)	0.014	(−0.010, 0.039)	23-23-1	0.36 (772), 0.29 (768), −0.15 (779)
x (xl-l)	0.013	(0.005, 0.022)	30-11-6	0.13 (783), 0.07 (772), −0.04 (786)
x5 (x5l-l)	0.012	(−0.011, 0.035)	23-21-3	0.38 (783), 0.17 (790), −0.13 (765)
	ΔP20			
e (xle-xl)	0.027	(−0.009, 0.063)	24-17-9	0.40 (768), 0.25 (800), −0.30 (792)
x (xl-l)	0.016	(−0.008, 0.040)	20-11-19	−0.25 (777), −0.15 (799), 0.20 (782)
x5 (x5l-l)	−0.015	(−0.063, 0.033)	18-19-13	0.40 (782), −0.30 (776), −0.35 (799)
	ΔHP20			
x (xl-l)	0.010	(−0.003, 0.022)	12-3-32	−0.15 (777), 0.10 (798), 0.10 (763)
e (xle-xl)	0.000	(−0.028, 0.028)	10-14-23	0.30 (768), 0.20 (799), −0.25 (770)
x5 (x5l-l)	−0.016	(−0.043, 0.011)	11-14-22	−0.35 (799), −0.25 (777), 0.20 (793)

humT05xle: The submitted humT05xle run was a blind feedback run based 50% on humT05xl and 25% each on expansion queries from the first 2 rows of humT05xl. (The expansion queries used a document length normalization of 750.) This run used the same approach as the humR05txle run described in the Robust section.

Table 9 shows that on the short title queries, the plain content search technique returned a relevant document in the first 10 rows for 94% of the topics (47/50). When available, a highly relevant was returned in the first 10 rows for 70% of the topics (33/47).

Note: even though the submitted runs included 10,000 rows per query, the mean scores in Table 9 are

Table 11: Mean Scores of Submitted Terabyte Efficiency Runs (20 Rows Retrieved)

Run	FRS	S@1	S@5	S@10	S@20	P20	MRR
humTE05i4ld	0.895	35/50	45/50	46/50	47/50	0.549	0.788
humTE05i5	0.807	28/50	38/50	41/50	46/50	0.446	0.650
humTE05i4	0.796	24/50	37/50	43/50	45/50	0.439	0.613
humTE05i4l	0.792	27/50	39/50	40/50	44/50	0.451	0.644
on adhoc titles							
(humTE05i4ld)	0.915	36/50	46/50	47/50	48/50	0.565	0.808
(humTE05i5)	0.813	29/50	38/50	41/50	46/50	0.454	0.666
(humTE05i4)	0.798	25/50	37/50	43/50	45/50	0.447	0.623
(humTE05i4l)	0.794	28/50	39/50	40/50	44/50	0.460	0.654

based on just the first 1000 rows.

- ‘x’ experiment: The modest proximity weighting technique (‘x’ experiment in Table 10) led to a statistically significant increase in MAP (and HMAP), though the individual impacts on average precision were small. It had bigger impacts on First Relevant Score, but was neutral on average. This result is similar to its impact in the Robust task.

- ‘x5’ experiment: For the heavier weight on proximity weighting (‘x5’ experiment in Table 10), the mean differences were not statistically significant. Compared to the modest proximity weight, there were larger per-topic impacts in each direction. Again, this result is fairly similar to its impact in the Robust task.

- ‘e’ experiment: The blind feedback technique (‘e’ experiment in Table 10) produced a statistically significant increase in MAP, but mean FRS was negatively impacted. Again, this result is fairly similar to its impact in the Robust task.

7.3 Efficiency Experiments

In the efficiency task, the titles of the 50 adhoc topics were seeded into a set of 50,000 web queries. The participants submitted the top-20 results for all 50,000 queries (only the results for the 50 adhoc topics were going to be judged, but it was not announced in advance which 50 queries those were). The efficiency task was held in the first week of July 2005; results were due before the adhoc topics were released.

Because the efficiency queries were released only one week before results were due, the average query time would have to be under 12 seconds to complete even one run. (We submitted the maximum of 4 runs.) The time constraints discouraged the use of performance-intensive techniques (such as query expansion from blind feedback) and encouraged investigation of the tradeoff of time and retrieval quality.

Compared to our submitted (adhoc) runs of last year, we were using a faster machine (2.8GHz), the documents were indexed in one table, the searches were conducted locally (not over a network share), and the index was re-organized so that term position information would not need to be read from disk if the search did not include term proximity constraints. Furthermore, in diagnostics on last year’s topics we found that retrieval quality for boolean-AND was similar to that for boolean-OR, particularly for early precision measures. Boolean-AND queries tend to be faster because they usually generate fewer internal matches (and hence involve less per-match processing such as relevance value calculations).

The techniques used for the 4 submitted runs of July 2005 are described below. The base technique was to use the SearchServer CONTAINS predicate to perform a boolean-AND of the words of the Title field of the topic.

humTE05i4: The submitted humTE05i4 run used a boolean-AND of the query words. Inflections from stemming were not enabled for this run (SET TERM_GENERATOR ‘’). Document length normalization

Table 12: Mean Scores of Submitted Terabyte Named Page Finding Runs

Run	FRS	S@1	S@5	S@10	S@1000	MRR
humTN05rdpl	0.457	67/252	109/252	117/252	200/252	0.335
humTN05dpl	0.488	76/252	115/252	125/252	201/252	0.371
humTN05pl	0.487	79/252	115/252	126/252	202/252	0.378
humTN05l	0.406	53/252	93/252	103/252	195/252	0.279

was not enabled (SET RELEVANCE_DLEN_IMP 0). The '2:4' relevance method was used (which squared the importance of inverse document frequency) because it gave higher reciprocal rank scores than '2:3' last year (though not significantly so). An example SearchSQL query is below. This run averaged 0.8 seconds per query, including fetching of the top-20 rows. (Note that the times are averaged over the 50,000 web queries which may have been longer on average than the typical adhoc query.)

```
SELECT RELEVANCE('2:4') AS REL, DOCNO
FROM GOV2
WHERE FT_TEXT CONTAINS 'Puerto'&'Rico'&'state'
ORDER BY REL DESC;
```

humTE05i4l: The submitted humTE05i4l run was the same as humTE05i4 except that linguistic expansion from English inflectional stemming was enabled (SET TERM_GENERATOR 'word!ftelp/inflect'). This run averaged 1.1 seconds per query (presumably the extra matches from inflections took more time to handle).

humTE05i4ld: The submitted humTE05i4ld run was the same as humTE05i4l except that document length normalization was enabled (SET RELEVANCE_DLEN_IMP 250). This run averaged 4.4 seconds per query, in part because in this implementation the 25 million document lengths were re-read from disk for each query.

humTE05i5: The submitted humTE05i5 run was the same as humTE05i4 except that the '2:5' relevance method was used instead of '2:4' which disabled the use of inverse document frequency. As expected, the average query time was the same as for humTE05i4.

Table 11 lists the mean scores of the submitted efficiency runs. Because at least 2 of the efficiency topics had spelling inconsistencies compared to the corresponding adhoc topics, we also list diagnostic runs performed using the same techniques on the titles of the adhoc topics. (Also note that because just the top-20 rows are included, FRS and MRR might be a little lower than they would be in the adhoc tables, which are evaluated on the top-1000 rows.)

The efficiency run which included document length normalization produced similar quality scores to the adhoc runs (e.g. Success@10 of 94%). (Diagnostic experiments on the terabyte collection were included in last year's paper [11].)

7.4 Named Page Finding Experiments

The Named Page Finding Task of the Terabyte Track was much like the Known-Item task of the Enterprise Track. For each of the 252 queries, the goal was to find the particular page. 187 of the queries just had one right answer; for the others, the extra right answers presumably were duplicates. Of course, the Terabyte task was searching a collection with more than 100 times as many documents as the Enterprise task, and the documents were from government web sites rather than an organization's emails.

The techniques used for the 4 submitted runs of August 2005 were as follows:

humTN05l: The submitted humTN05l run was a plain content search including linguistic expansion from English inflectional stemming. This run used the same approach as the humEK05l run described in the Enterprise section (which has example query syntax).

Table 13: Impact of Terabyte Named Page Finding Techniques

Expt	ΔFRS	95% Conf	vs.	3 Extreme Diffs (Topic)
p (pl-l)	0.082	(0.054, 0.109)	106-27-119	1.00 (658), 1.00 (686), −0.40 (621)
d (dpl-pl)	0.001	(−0.010, 0.011)	37-52-163	−1.00 (658), 0.29 (820), 0.34 (672)
r (rdpl-dpl)	−0.031	(−0.047,−0.015)	16-83-153	−0.79 (686), −0.71 (667), 0.63 (658)
	ΔS1			
p (pl-l)	0.103	(0.063, 0.144)	27-1-224	1.00 (849), 1.00 (602), −1.00 (695)
d (dpl-pl)	−0.012	(−0.030, 0.006)	1-4-247	−1.00 (632), −1.00 (658), 1.00 (657)
r (rdpl-dpl)	−0.036	(−0.067,−0.005)	3-12-237	−1.00 (675), −1.00 (615), 1.00 (632)
	ΔS10			
p (pl-l)	0.091	(0.054, 0.128)	23-0-229	1.00 (839), 1.00 (838), 0.00 (872)
d (dpl-pl)	−0.004	(−0.018, 0.010)	1-2-249	−1.00 (814), −1.00 (658), 1.00 (641)
r (rdpl-dpl)	−0.032	(−0.057,−0.006)	1-9-242	−1.00 (641), −1.00 (817), 1.00 (658)
	ΔMRR			
p (pl-l)	0.099	(0.067, 0.131)	106-27-119	1.00 (686), 1.00 (658), −0.50 (695)
d (dpl-pl)	−0.008	(−0.020, 0.005)	37-52-163	−1.00 (658), −0.75 (686), 0.50 (657)
r (rdpl-dpl)	−0.035	(−0.057,−0.014)	16-83-153	−0.94 (667), −0.91 (849), 0.67 (676)

humTN05pl: The submitted humTN05pl run was the same as humTN05l except that it put an additional 10% weight on matches in the ALL_PROPS column (which included the title, meta tags, url, etc.) and an additional 10% weight on phrase matches in ALL_PROPS. This run used the same approach as the humEK05pl run described in the Enterprise section (which has example query syntax).

humTN05dpl: The submitted humTN05dpl run was the same as humTN05pl except that it put additional weight on urls of depth 4 or less. This run used the same approach as last year's humW04dpl run (for which example syntax is given in [11]) except document length normalization (RELEVANCE_DLEN_IMP) was set to 250 instead of 500.

humTN05rdpl: The submitted humTN05rdpl run was the same as humTN05dpl except that it put additional weight on the url type. This run used the same approach as last year's humW04rdpl run (for which example syntax is given in [11]) except document length normalization (RELEVANCE_DLEN_IMP) was set to 250 instead of 500.

Table 12 lists the mean scores of the 4 submitted runs. The plain content search had a Success@10 of just 41% (103/252), and adding more weight on other columns boosted Success@10 to just 50% (126/252). These success rates are lower than in the named page finding subtask of last year's Web Track (for which the corresponding scores were 65% and 76%). Perhaps terabyte named page finding is harder because the GOV2 collection has 20 times as many pages as the GOV collection of last year's Web Track (a bigger haystack in which to find a needle).

- 'p' experiment: Table 13 shows that the 'p' factor (extra weight on columns such as the Title) led to statistically significant increases in mean FRS, S@1, S@10 and RR. This result is similar to its effect in this year's Enterprise Known-Item task and on last year's Web Named Page queries.

- 'd' experiment: Table 13 shows that the 'd' factor (modest extra weight for less deep urls) was of neutral impact on average, though on some topics it had a substantial impact in each direction. This result is similar to that for last year's Web Named Page queries.

- 'r' experiment: Table 13 shows that the 'r' factor (strong extra weight for urls of root, subroot or path types) led to statistically significant decreases in mean FRS, S@1, S@10 and RR. This result is similar to that for last year's Web Named Page queries.

8 Conclusions

The First Relevant Score measure (FRS) was successful at detecting the impact of various retrieval techniques on the first relevant item retrieved.

- In the known-item tasks, FRS found a statistically significant mean difference for 7 of the 8 experiments, more than for MRR (6/8), S@1 (6/8) and S@10 (4/8). Also, the largest per-topic differences for FRS were not skewed to minor differences in the early ranks, a common shortcoming of reciprocal rank.
- In the ad hoc tasks, FRS found a statistically significant mean difference for 11 of the 22 experiments, almost as many as for MAP (12/22). Unlike for MAP, the differences for FRS are known to apply to the first relevant item retrieved.

The various retrieval measures moved together for most retrieval techniques. The one case for which we saw retrieval measures move significantly in opposite directions was for the "blind feedback" technique, which boosted secondary recall measures (such as MAP) but was detrimental to primary recall measures (such as FRS).

(In principle, duplicate filtering should have the opposite effect to blind feedback; i.e. successful duplicate filtering techniques would potentially increase primary recall measures (by filtering out duplicate non-relevants before the first desired item) but may decrease secondary recall measures (which may require duplicate relevants to be retrieved to get a maximum score). However, we did not experiment with duplicate filtering in this paper.)

We did not find cases where restricting to highly relevants or restricting to judged documents made a substantial difference to the rating of the techniques.

We did not find cases where secondary recall measures disagreed significantly with each other. In particular, GMAP still moved with MAP, even for the non-robust blind feedback technique. Precision@20 also moved with MAP.

So if one wants to contrast MAP with another measure for ad hoc tasks, it appears that one should use a primary recall measure, such as FRS.

References

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

[2] Nick Craswell. W3C Test Collection. http://research.microsoft.com/users/nickcr/w3c-summary.html

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

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

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

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

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

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

[9] Stephen Tomlinson. Experiments in Named Page Finding and Arabic Retrieval with Hummingbird SearchServer^{TM} at TREC 2002. Proceedings of TREC 2002.

[10] Stephen Tomlinson. Robust, Web and Genomic Retrieval with Hummingbird SearchServer^{TM} at TREC 2003. Proceedings of TREC 2003.

[11] Stephen Tomlinson. Robust, Web and Terabyte Retrieval with Hummingbird SearchServer^{TM} at TREC 2004. Proceedings of TREC 2004.

[12] TREC-2005 Enterprise Track Guidelines. http://www.ins.cwi.nl/projects/trec-ent/wiki/index.php/Main_Page

[13] TREC 2005 Terabyte Track Guidelines. http://plg.uwaterloo.ca/∼claclark/TB05.html

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

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