Robust, Web and Genomic Retrieval with Hummingbird SearchServer™ at TREC 2003

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

February 4, 2004

Abstract

Hummingbird participated in 4 tasks of TREC 2003: the ad hoc task of the Robust Retrieval Track (find at least one relevant document in the first 10 rows from 1.9GB of news and government data), the navigational task of the Web Track (find the home or named page in 1.2 million pages (18GB) from the .GOV domain), the topic distillation task of the Web Track (find key resources for topics in the first 10 rows from home pages of .GOV), and the primary task of the Genomics Track (find all records focusing on the named gene in 1.1GB of MEDLINE data). In the ad hoc task, SearchServer found a relevant document in the first 10 rows for 48 of the 50 new short (Title-only) topics. In the navigational task, SearchServer returned the home or named page in the first 10 rows for more than 75% of the 300 queries. In the distillation task, a SearchServer run found the most key resources in the first 10 rows of the submitted runs from 23 groups.

1 Introduction

Hummingbird SearchServer[1] is an indexing, search and retrieval engine for embedding in Windows and UNIX information applications. SearchServer, originally a product of Fulcrum Technologies, was acquired by Hummingbird in 1999. Founded in 1983 in Ottawa, Canada, Fulcrum produced the first commercial application program interface (API) for writing information retrieval applications, Fulcrum® Ful/Text™. The SearchServer kernel is embedded in many Hummingbird products, including SearchServer, an application toolkit used for knowledge-intensive applications that require fast access to unstructured information.

SearchServer supports a variation of the Structured Query Language (SQL), SearchSQL™, which has extensions for text retrieval. SearchServer conforms to subsets of the Open Database Connectivity (ODBC) interface for C programming language applications and the Java Database Connectivity (JDBC) interface for Java applications. Almost 200 document formats are supported, such as Word, WordPerfect, Excel, PowerPoint, PDF and HTML.

SearchServer works in Unicode internally [5] and supports most of the world's major character sets and languages. The major conferences in text retrieval evaluation (TREC [9], CLEF [1] and NTCIR [7]) have provided opportunities to objectively evaluate SearchServer's support for more than a dozen languages.

This paper looks at experimental work with SearchServer for robust retrieval (robustness of ad hoc search across topics), web navigation (find the one page the user wanted, i.e. a known-item search task), web distillation (find key resource pages for broad topics), and genomic retrieval (a domain-specific task). For the submitted runs in August 2003, an experimental post-5.x development build of SearchServer was used.

[1]Fulcrum® is a registered trademark, and SearchServer™, SearchSQL™, Intuitive Searching™ and Ful/Text™ are trademarks of Hummingbird Ltd. All other copyrights, trademarks and tradenames are the property of their respective owners.

2 Robust Retrieval

The document set of the TREC 2003 Robust Retrieval Track was a subset of the news and government data of TREC Disks 4 and 5. It consisted of 528,155 documents totaling 1,997,002,586 bytes (1.9 GB). The average document size was 3781 bytes. For more information, see the track overview paper.

For this ad hoc task, participants were asked to focus not just on mean average precision but on at least one other measure indicative of "robustness" across results, such as the number of topics for which at least one relevant was retrieved in the first 10 rows.

2.1 Indexing

The custom text reader called cTREC, described in our first TREC paper [10], already supported detailed handling of the TREC Disk collections. For example, it allowed indexing of text following particular tags (such as <HEADLINE> and <TEXT>) and disabled indexing for text surrounded by other tags (such as <PAGE>...</PAGE>) and for the tags themselves. As this year's guidelines did not restrict the fields allowed for indexing, we used the /k option of cTREC to allow indexing of text tagged as keywords (in particular, text tagged by <IN> or <SUBJECT> in the case of the disks used this year). Past experiments suggest that this detailed handling does not affect the results much.

We used the mygov.stp stopword list (99 English stopwords) first used for a web task last year [12]. The option to support inflections from lexical English stemming was enabled. We also experimented with an option to construct term vectors for result list clustering for this task.

2.2 Searching

The submitted humR03d run was a "plain" SearchServer run on the Description field of each topic. It used SearchServer's Intuitive Searching (i.e. the IS_ABOUT predicate of SearchSQL). Here is an example SearchSQL query for topic 314:

```
SELECT RELEVANCE('V2:3') AS REL, DOCNO
FROM ROBUST03
WHERE FT_TEXT IS_ABOUT 'Commercial harvesting of marine vegetation
  such as algae, seaweed and kelp for food and drug purposes.'
ORDER BY REL DESC;
```

SearchServer's relevance value calculation is the same as described last year [12]. Briefly, SearchServer dampens the term frequency and adjusts for document length in a manner similar to Okapi [8] and dampens the inverse document frequency using an approximation of the logarithm. SearchServer's relevance values are always an integer in the range 0 to 1000.

Before the queries were run, various SET statements were issued. "SET MAX_SEARCH_ROWS 1000" ensured the resulting working table would contain at most 1000 rows. Inflections from English stemming were enabled by "SET VECTOR_GENERATOR 'word!ftelp/lang=english/base/noalt | * | word!ftelp/lang=english/inflect' " (for more details on stemming for several European languages, see our CLEF paper [14]). The importance of document length to the relevance value calculation was set with "SET RELEVANCE_DLEN_IMP 750" (scale of 0 to 1000).

We automatically removed "query stop words" such as "find", "relevant" and "document" from the topics before presenting them to SearchServer, i.e. words which are not stop words in general but were commonly used in previous years' TREC and CLEF topics as general instructions (this year's topics were not reviewed). An evaluation in last year's CLEF paper [11] found this step to be of only minor impact in several European languages including English.

The submitted humR03t run was the same as humR03d except that the Title field of the topic was used instead of the Description. For example, for topic 314, the Where clause was just "WHERE FT_TEXT IS_ABOUT 'Marine Vegetation' ". This run represented a "plain" SearchServer run for Title queries.

The submitted humR03de run used query expansion from blind feedback. The first two rows of the humR03d run were used to find additional query terms. Only terms appearing in at most 5% of the documents

(based on the most common inflection of the term) were included. Mathematically, the approach is similar to Rocchio feedback with weights of one-half for the original query and one-quarter for each of the 2 expansion rows. This was the same blind feedback approach as used for Arabic experiments at TREC last year [12] except that we just used 2 rows for expansion this time instead of 5 (diagnostics on past CLEF and TREC ad hoc tasks suggested fewer rows may be more effective, perhaps because most tasks have fewer relevant documents to feed back in per topic than the Arabic task did). Blind feedback from the top retrieved documents is often effective at increasing recall later in the result list without sacrificing early precision, important components of the average precision measure. However, the large number of additional query terms negatively impacts performance. In practice, users could manually add terms to the query rather than work blindly. For this run, the measure in mind was average precision.

The submitted humR03dc run re-ordered the top 100 rows of humR03d so that the first 10 rows were from different clusters to see if that increased the chance of a relevant document in the top 10 rows. The steps were as follows:

First, the parameter settings were the same as for humR03d except that "SET MAX_SEARCH_ROWS 100" was used instead of 1000. Hence a relevant document had to appear in the top 100 for re-ordering to have a chance of moving it into the top 10.

Next, a result list clustering query was run on the 100 row working table, producing a set of up to 10 clusters. Each of the 100 rows appeared in exactly one cluster; the number of rows in each cluster could differ. Within each cluster, the rows were ordered by the original relevance value. The clusters themselves were ordered by the average relevance value of the rows of the cluster. The clustering was based on the term vectors of the documents built at index-time. Other than its impact on which 100 documents were clustered, the query had no impact on the clustering.

Finally, a round-robin of the clusters was followed, with the row of highest remaining relevance score of each cluster placed into the final result. (In the submitted file, the first 3 digits after the decimal point are the relevance value of the document, and the 4th digit is the cluster number from 0-9.)

The top 100 rows of humR03d and humR03dc should be the same except for the order. The first row for a topic in humR03d would appear somewhere in the top 10 for the topic in humR03dc. The other rows in the top 10 might differ.

The final result of humR03dc had at most 100 rows per topic. We did not bother to pad to 1000 rows. Hence for this run there was a bias against measures which consider documents retrieved past 100 rows, such as recall and average precision. For this run, the measure in mind was the 'relevant in the top 10 rows' measure.

The submitted humR03tc run was the same as humR03dc except that it was based on humR03t instead of humR03d.

2.3 Results

For this task, there were 50 "old" topics and 50 "new" topics.

The 50 old topics were selected (by the task organizers) from past years' ad hoc TREC topics 301-450 to produce a set of "tough" topics, i.e. topics on which few systems produced a high precision score when they were originally used (though there may have been a bias against topics on which all systems produced a low score; the track overview paper may elaborate more). As they were already in the public domain, the guidelines allowed groups to continue to study these topics for this year's submissions, which might also help lead to techniques for improving results on tough topics. One must be cautious however at reading too much into results on these topics (even "statistically significant" results) because of the possibility that the techniques are tuned to this data.

For the 50 new topics, (automatic) systems were not allowed to be altered based on examination of the topics, so in that sense the results may be more meaningful. But there was no reason to expect these topics to be as "tough" as the specially-selected older set (it is very hard to predict which topics will be tough in advance) so for the purpose of this track (robustness across topics) there may not be enough challenging topics to distinguish the techniques.

We separately list the results for each of these sets of topics (we do not bother to look at the combined scores). Also, for the 50 new topics, the relevance assessors distinguished "highly relevant" documents from

Table 1: Precision of Submitted Runs

Run	AvgP	P@5	P@10	P@20	Rec0	Rec30	P@R	%Rel10
(humR03te-old)	0.131	34.4%	33.0%	27.7%	0.564	0.189	19.4%	40/50
humR03t-old	0.109	30.4%	28.8%	23.9%	0.555	0.151	17.1%	42/50
humR03tc-old	0.057	20.8%	17.8%	18.1%	0.476	0.070	12.6%	38/50
humR03de-old	0.148	38.8%	35.2%	28.1%	0.656	0.187	19.6%	38/50
humR03d-old	0.127	37.2%	29.6%	24.7%	0.635	0.167	17.4%	39/50
humR03dc-old	0.071	26.0%	20.6%	17.3%	0.582	0.072	13.4%	41/50
(humR03te-new)	0.332	51.2%	44.8%	35.8%	0.684	0.482	33.6%	46/50
humR03t-new	0.280	46.8%	41.0%	32.2%	0.672	0.401	30.6%	48/50
humR03tc-new	0.147	28.4%	20.4%	19.7%	0.630	0.194	18.9%	42/50
humR03de-new	0.377	57.2%	48.4%	39.9%	0.767	0.543	37.7%	43/50
humR03d-new	0.346	57.6%	47.6%	38.6%	0.779	0.500	34.5%	46/50
humR03dc-new	0.178	33.6%	23.4%	21.1%	0.687	0.227	20.8%	44/50
(humR03te-newH)	0.253	28.8%	21.2%	16.2%	0.430	0.342	25.4%	31/43
humR03t-newH	0.225	24.6%	19.8%	15.1%	0.402	0.307	24.1%	31/43
humR03tc-newH	0.117	9.8%	8.1%	8.0%	0.333	0.140	10.0%	22/43
humR03de-newH	0.309	30.2%	24.2%	17.8%	0.551	0.424	28.2%	32/43
humR03d-newH	0.305	32.6%	23.5%	17.2%	0.579	0.413	29.8%	31/43
humR03dc-newH	0.176	18.6%	11.6%	10.2%	0.491	0.225	17.4%	27/43

Table 2: Impact of Clustering on Percentage of Topics With a Relevant in Top 10

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
D-old-Rel10	0.040	(−0.101, 0.181)	7-5-38	1.000 (330), 1.000 (401)
D-new-Rel10	−0.040	(−0.121, 0.041)	1-3-46	1.000 (605), −1.000 (608)
T-old-Rel10	−0.080	(−0.161,−0.019)	0-4-46	−1.000 (394), −1.000 (336)
D-newH-Rel10	−0.093	(−0.210, 0.001)	1-5-37	1.000 (643), −1.000 (616)
T-new-Rel10	−0.120	(−0.221,−0.039)	0-6-44	−1.000 (612), −1.000 (642)
T-newH-Rel10	−0.209	(−0.373,−0.069)	2-11-30	1.000 (631), 1.000 (620)

just "relevant" documents. 43 of the 50 topics had at least one "highly relevant" document. We list the scores averaged over those 43 topics when just considering highly relevants as relevant (tagged with "newH").

Table 1 gives an overview of several precision scores for each submitted run (also, in brackets, is an unsubmitted run (because of the 5-run submission limit) produced at the same time, humR03te, an analog of humR03de for Titles). Listed for each run are its mean average precision (AvgP), the mean precision after

Table 3: Impact of Blind Feedback on Average Precision

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
(T-new-AvgP)	0.052	(0.031, 0.075)	40-10-0	0.345 (614), 0.188 (607)
D-new-AvgP	0.031	(0.004, 0.057)	34-16-0	−0.221 (616), 0.205 (633)
(T-newH-AvgP)	0.027	(0.012, 0.043)	30-10-3	0.180 (648), 0.147 (626)
(T-old-AvgP)	0.022	(0.010, 0.035)	32-18-0	0.181 (350), 0.174 (372)
D-old-AvgP	0.021	(0.008, 0.035)	32-17-1	0.167 (350), 0.146 (320)
D-newH-AvgP	0.004	(−0.016, 0.022)	25-17-1	−0.213 (644), −0.128 (614)

Table 4: Impact of Blind Feedback on Percentage of Topics With a Relevant in Top 10

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
D-newH-Rel10	0.023	(−0.070, 0.117)	3-2-38	1.000 (601), 1.000 (633)
(T-newH-Rel10)	0.000	(−0.094, 0.094)	2-2-39	1.000 (636), 1.000 (628)
D-old-Rel10	−0.020	(−0.081, 0.041)	1-2-47	1.000 (356), −1.000 (426)
(T-old-Rel10)	−0.040	(−0.141, 0.061)	2-4-44	1.000 (356), 1.000 (439)
(T-new-Rel10)	−0.040	(−0.121, 0.041)	1-3-46	1.000 (627), −1.000 (632)
D-new-Rel10	−0.060	(−0.141, 0.001)	0-3-47	−1.000 (632), −1.000 (610)

5, 10 and 20 documents retrieved (P@5, P@10 and P@20 respectively), the mean interpolated precision at 0% and 30% recall (Rec0 and Rec30 respectively), the mean precision after R documents retrieved (P@R) where R is the number of relevant documents for the topic, and the ratio of the number of topics with at least one relevant retrieved in the top 10 vs. the total number of topics (%Rel10). (Definitions of the measures are in last year's paper [12], and they likely are also in an appendix of the conference proceedings.) It appears that for every measure listed, the score on the "new" topics is higher than the corresponding score for the "old" topics, i.e. as expected, the "old" topics were more challenging (on average).

For tables focusing on the impact of one particular difference in approach, the columns are as follows:

- "Experiment" indicates whether the Title or Description topics were used ("T" or "D" respectively) and whether the score is based on the old topics ("old"), the new topics when treating all relevants the same ("new"), or the new topics just counting highly relevants as relevant ("newH").
- "AvgDiff" is the average (mean) difference in the score.
- "95% Confidence" is an approximate 95% confidence interval for the average difference calculated using Efron's bootstrap percentile method[2] [3] (using 100,000 iterations). 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, 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 score was higher, lower and tied (respectively) with the feature enabled. These numbers should always add to the number of topics (50 or 43).
- "2 Largest Diffs (Topic)" lists the two largest differences in the score (based on the absolute value) with each followed by the corresponding topic number in brackets (the old topic numbers range from 301 to 450 and the new topic numbers from 601 to 650).

Table 2 shows the impact of the clustering-based technique on the percentage of topics with a relevant in the first 10 rows. For Description queries, this is based on subtracting the scores of humR03d from humR03dc, and for the Title queries, subtracting the scores of humR03t from humR03tc. As you can see, there was a net gain of 2 topics with a relevant in the top 10 on the old Description queries (7 gained but 5 lost), though this was not statistically significant, and the loss of 4 on the old Title queries was statistically significant. On the new queries, the finding was similar; the differences were not significant on the Description queries but were on the Title queries, both when counting all relevants or just highly relevants.

Table 3 shows the impact of the blind feedback technique on the average precision score (based on subtracting humR03d from humR03de, and humR03t from (unsubmitted run) humR03te). The increase was statistically significant for 5 of the 6 cases, the exception being for highly relevants on the new Description topics.

Table 4 shows the impact of the same blind feedback technique on the perentage of topics with a relevant in the first 10 rows (based on the same runs as Table 3). None of the impacts were statistically significant.

[2]See [11] for some comparisons of confidence intervals from the bootstrap percentile, Wilcoxon signed rank and standard error methods for both average precision and Precision@10.

Table 5: Examples of URL type and depth values

URL	Type	Depth	Depth Term
http://nasa.gov/	ROOT	1	URLDEPTHA
http://www.nasa.gov/	ROOT	1	URLDEPTHA
http://jpl.nasa.gov/	ROOT	2	URLDEPTHAB
http://fred.jpl.nasa.gov/	ROOT	3	URLDEPTHABC
http://nasa.gov/jpl/	SUBROOT	2	URLDEPTHAB
http://nasa.gov/jpl/fred/	PATH	3	URLDEPTHABC
http://nasa.gov/index.html	ROOT	1	URLDEPTHA
http://nasa.gov/fred.html	FILE	2	URLDEPTHAB

Table 6: Number of Pages of each URL Type and Depth

Type	#Pages	Depth	#Pages	Depth	#Pages
ROOT	6,906	1	635	6	269,949
SUBROOT	18,179	2	16,792	7	136,513
PATH	55,332	3	128,898	8	44,960
FILE	1,167,336	4	282,086	9	15,289
		5	344,694	10+	7,937

For the plain SearchServer runs, more topics found a relevant in the first 10 rows using the Titles than the Descriptions (42 to 39 for the old topics, 48 to 46 for the new topics (though tied at 31 when restricting to highly relevants) as per Table 1). These results might suggest that shorter queries are more "robust" (perhaps extra details throw off a system more often than missing details, even though the longer queries score higher on the other listed measures which reward recall more). However, these changes in the number of topics with a relevant in the first 10 rows did not pass a significance test.

3 Web Retrieval

Both tasks of the TREC 2003 Web Track used the same .GOV collection as last year. It consists of pages downloaded from the .gov domain of the World Wide Web in early 2002. Uncompressed, it is 19,455,030,550 bytes (18.1 GB) and a total of 1,247,753 documents. The average document size is 15,592 bytes. For more information on the .GOV collection, see [4].

3.1 Indexing

The indexing approach was the same as described in last year's paper [12] (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, the custom text reader cTREC populated particular columns such as TITLE (if any), URL, URL_TYPE and URL_DEPTH. 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. Table 5 contains URL types and depths for example URLs, and Table 6 shows the number of .GOV pages of each URL type and depth. The exact rules we used are given in last year's paper [12].

3.2 Searching

Even though the 2 web tasks are potentially quite different (the navigational task is a known-item task (one right answer), while the topic distillation task is focused on distilling broad topics to key resource pages), we used the same techniques for both tasks for each of the 5 submitted runs (and most of the techniques used were the same as last year). This allows us to compare the impact of the techniques on different tasks.

The submitted humNP03l and humTD03l runs used the same approach as the diagnostic base run described in last year's paper [12] which was just to search the content (FT_TEXT column) using the IS_ABOUT predicate (i.e. the same approach as used for the "plain" runs of the Robust task).

The submitted humNP03pl and humTD03pl runs used the same approach as last year's hum02pd run. Below is an example SearchSQL query. The queries differed from humNP03l and humTD03l in that that properties and phrases in properties were given a little extra weight. (The ALL_PROPS column contained the title, URL, first heading and some meta tags, but not most of the document content; see last year's paper for the details.) Note that the FT_TEXT column also indexed all of the properties except for the URL.

```
SELECT RELEVANCE('V2:3') AS REL, DOCNO
FROM GOV
WHERE
 (ALL_PROPS CONTAINS 'visiting pandas national zoo' WEIGHT 1) OR
 (ALL_PROPS IS_ABOUT 'visiting pandas national zoo' WEIGHT 1) OR
 (FT_TEXT IS_ABOUT 'visiting pandas national zoo' WEIGHT 10)
ORDER BY REL DESC;
```

The CONTAINS predicate does phrase searching, so the listed terms would have to occur adjacently in the specified order (except stop words). "SET PHRASE_DISTANCE 4" was previously specified so that there could be up to 4 characters between adjacent terms (plus additional whitespace). By default, the CONTAINS predicate does exact searching (i.e. no inflections from stemming), though some Unicode-based normalizations (e.g. decompositions and conversion to upper-case) are still done. The motivation for including the query as a phrase was that it seemed the query might often be in the title or other property information of the document (e.g. a query in mind was "Washington State Legislature" (which was not one of the 150 official queries last year)). The phrase searching was just given one-tenth the weight of content searching for relevance ranking purposes. Experiments on the TREC 2001 entry page finding task suggested a small weight was helpful (on average) but a strong weight had a negative impact.

The IS_ABOUT predicate uses SearchServer's Intuitive Searching. It by default matches inflections froms English stemming and just requires one of the terms to have a match. It was used with WEIGHT 1 on the ALL_PROPS column to increase the ranking of documents with query terms in the title or other property information. It was used with WEIGHT 10 on the FT_TEXT column (which represents the external document). Again, these weights were chosen based on what worked well on the TREC 2001 entry page finding task.

The submitted humNP03upl and humTD03upl runs used the same approach as last year's hum02upd run. The 'u' indicates a higher weight was given to URLs of particular type and depth. See last year's paper for an example of the SearchSQL syntax [12].

The submitted humNP03uhpl and humTDuhpl runs used the same approach as last year's hum02uhp run except for using a document length importance of 500 instead of 250 (500 was used for all submitted web runs this year). The 'h' indicates an even higher weight was given to URL_TYPE (the 3 terms of WEIGHT 10 were given WEIGHT 25). On the TREC 2001 entry page finding task, the stronger URL_TYPE weights gave similar MRR scores to the lower ones.

The submitted humNP03up and humTD03up runs were the same as humNP03upl and humTD03upl (respectively) except that linguistic expansion from English stemming was disabled (i.e. matching of inflections was disabled) by "SET VECTOR_GENERATOR '' ".

For the navigational (humNP03*) runs, the statement "SET MAX_SEARCH_ROWS 50" was previously executed so that the working table would contain at most 50 rows, whereas for the topic distillation (humTD03*) runs, the statement "SET MAX_SEARCH_ROWS 1000" was previously executed.

Table 7: Scores of Submitted Navigational Runs

Run	300 MRR	300 %Top10	300 %Fail	HP 150 MRR	HP 150 %Top10	HP 150 %Fail	NP 150 MRR	NP 150 %Top10	NP 150 %Fail
humNP03up	0.545	77.3%	12.3%	0.584	82.0%	8.7%	0.506	72.7%	16.0%
humNP03upl	0.535	77.7%	11.7%	0.591	82.7%	8.0%	0.480	72.7%	15.3%
humNP03pl	0.465	68.3%	17.3%	0.361	56.7%	27.3%	0.568	80.0%	7.3%
humNP03uhpl	0.386	56.7%	26.0%	0.500	70.7%	16.7%	0.271	42.7%	35.3%
humNP03l	0.321	54.3%	25.7%	0.223	41.3%	40.7%	0.420	67.3%	10.7%

Table 8: Impact of Submitted Navigational Techniques on Reciprocal Rank

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
HP u (upl - pl)	0.229	(0.168, 0.291)	87-17-46	1.000 (321), 1.000 (395)
HP p (pl - l)	0.139	(0.093, 0.187)	72-14-64	1.000 (422), 1.000 (200)
HP l (upl - up)	0.007	(−0.011, 0.026)	14-13-123	0.667 (271), 0.500 (349)
HP h (uhpl - upl)	−0.091	(−0.134,−0.050)	14-57-79	−1.000 (355), −0.977 (300)
NP u (upl - pl)	−0.088	(−0.134,−0.042)	13-66-71	−1.000 (359), 0.950 (154)
NP p (pl - l)	0.147	(0.092, 0.204)	74-17-59	−1.000 (416), 0.975 (151)
NP l (upl - up)	−0.026	(−0.055, 0.002)	14-32-104	0.889 (215), −0.875 (306)
NP h (uhpl - upl)	−0.209	(−0.257,−0.162)	1-92-57	−1.000 (249), −1.000 (154)

For the web queries, no query terms were discarded (e.g. there was no expectation that discarding the words "find", "relevant" and "document" would be beneficial, unlike for the Robust task). Of course, the index omitted a few stop words (e.g. "the", "by") as previously mentioned.

SearchServer's relevance value calculation is the same as described for the Robust task. Additionally, when multiple predicates are combined, as was done for some of the web approaches, SearchServer currently does not normalize by query length. For example, the URL_TYPE clauses would have a lot less relative impact if the topic query contained 5 words instead of 1.

3.3 Results

The evaluation measures are likely explained in an appendix of this volume. Briefly, for the navigational task, "Reciprocal Rank" for a topic is one divided by the rank in which the home or named page was found (using the smallest rank if there were duplicates of the page), or zero if the page was not found. "Mean Reciprocal Rank" (MRR) is the average of the reciprocal ranks over all the topics. "%Top10" is the percentage of topics for which the home or named page was found in the first 10 rows. "%Fail" is the percentage of topics for which the home or named page was not found in the first 50 rows. The topic distillation measures are the same as described previously in the Robust section.

Table 7 shows the scores of the submitted navigational runs in descending order by mean reciprocal rank over all 300 queries. The HP columns show the scores just for the 150 home page queries. The NP columns show the scores just for the 150 named page queries. (The topics did not state whether they were of HP or NP type; that information was provided by the organizers after the submission date for use in analysis.)

Table 8 shows the impact when isolating each technique distinguishing the submitted navigational runs:

- The 'u' factor (extra weight for URL type and depth) increased MRR dramatically on the home pages (23 points) but (like last year) was detrimental on the named pages (9 points). More diagnostics are below.
- The 'p' factor (extra weight for HTML properties and phrases in properties) increased MRR 14 points on both home and named pages. More diagnostics are below.

Table 9: Diagnostics of Extra Weight on Document Structure (Navigational Task, Reciprocal Rank)

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
HP v (vl-l)	0.079	(0.045, 0.115)	60-19-71	0.933 (425), 0.917 (333)
HP q (ql-l)	0.056	(0.027, 0.088)	44-15-91	0.976 (385), 0.909 (307)
HP qv (vql-l)	0.115	(0.073, 0.159)	62-19-69	1.000 (334), 1.000 (389)
HP v (vql-ql)	0.058	(0.025, 0.095)	44-27-79	0.978 (389), 0.917 (425)
HP q (vql-vl)	0.036	(0.009, 0.066)	27-18-105	1.000 (389), 0.800 (322)
HP other (pl-vql)	0.024	(−0.015, 0.064)	37-29-84	1.000 (266), −0.857 (334)
NP v (vl-l)	0.120	(0.073, 0.169)	74-16-60	0.975 (151), 0.969 (286)
NP q (ql-l)	0.050	(0.014, 0.089)	41-20-89	−1.000 (416), 0.975 (151)
NP qv (vql-l)	0.128	(0.078, 0.180)	71-18-61	−1.000 (416), 0.975 (151)
NP v (vql-ql)	0.078	(0.041, 0.117)	55-17-78	0.963 (178), −0.909 (304)
NP q (vql-vl)	0.008	(−0.017, 0.032)	21-18-111	−0.857 (248), −0.750 (259)
NP other (pl-vql)	0.019	(−0.013, 0.051)	30-27-93	−0.938 (196), −0.800 (449)

Table 10: Diagnostics of Extra Weight on URL Structure (Navigational Task, Reciprocal Rank)

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
HP d5 (d5pl-pl)	0.158	(0.110, 0.207)	73-7-70	1.000 (384), 1.000 (395)
HP d10 (d10pl-pl)	0.205	(0.151, 0.261)	85-10-55	1.000 (244), 1.000 (384)
HP d15 (d15pl-pl)	0.213	(0.152, 0.275)	84-16-50	1.000 (184), 1.000 (244)
HP d20 (d20pl-pl)	0.173	(0.109, 0.238)	79-22-49	1.000 (392), 1.000 (395)
HP r5 (r5pl-pl)	0.168	(0.121, 0.217)	76-8-66	0.941 (201), 0.941 (328)
HP r10 (r10pl-pl)	0.213	(0.157, 0.269)	81-13-56	1.000 (395), 1.000 (321)
HP r5d5 (r5d5pl-pl)	0.231	(0.176, 0.288)	87-11-52	1.000 (184), 1.000 (321)
NP d5 (d5pl-pl)	−0.001	(−0.024, 0.024)	23-29-98	0.950 (154), 0.750 (264)
NP d10 (d10pl-pl)	−0.027	(−0.062, 0.008)	20-45-85	−0.952 (359), 0.950 (154)
NP d15 (d15pl-pl)	−0.084	(−0.127,−0.043)	16-61-73	−1.000 (359), −0.875 (189)
NP d20 (d20pl-pl)	−0.166	(−0.215,−0.119)	10-79-61	−1.000 (359), −1.000 (249)
NP r5 (r5pl-pl)	−0.013	(−0.039, 0.013)	11-34-105	0.833 (383), 0.750 (264)
NP r10 (r10pl-pl)	−0.069	(−0.109,−0.030)	10-61-79	−0.941 (359), −0.900 (216)
NP r5d5 (r5d5pl-pl)	−0.040	(−0.075,−0.005)	19-48-83	0.950 (154), −0.929 (359)

- The 'l' factor (linguistic expansion (inflections) from lexical English stemming) made little difference (on average).
- The 'h' factor (even more extra weight for URL type) was detrimental even on the home page queries, even though it had a neutral impact on the TREC 2001 entry page task.

Table 9 isolates the components of the 'p' factor. 'v' denotes that the run included TITLE IS_ABOUT (i.e. vector) matching with weight 1, and 'q' denotes that the run included TITLE CONTAINS (i.e. phrase) matching with weight 1. Adding the 'v' factor (to a full content search with weight 10) increased MRR significantly for both home pages and named pages (8 and 12 points respectively as per the "v (vl-l)" rows). The 'q' factor had significant, though smaller, increases (6 and 5 points respectively as per the "q (ql-l)" rows). If one of these was already done, adding the other still led to a significant increase except in the case of adding phrase matching to a vector match for named pages (as per the "NP q (vql-vl)" row). Using the ALL_PROPS column instead of the TITLE column did not lead to a further significant increase as per "other (pl-vql)" rows. So for the 'p' factor, like last year, most of the benefit appears to have come from the TITLE weighting, but unlike last year, both vector and phrase matching helped significantly, not just vector

Table 11: Scores of Submitted Topic Distillation Runs

Run	AvgP	P@5	P@10	P@20	Rec0	Rec30	P@R	Topics
humTD03upl	0.139	14.4%	12.8%	9.2%	0.382	0.166	14.9%	50
humTD03up	0.120	14.8%	12.4%	8.9%	0.362	0.147	13.3%	50
humTD03uhpl	0.098	13.2%	10.2%	6.9%	0.357	0.105	10.7%	50
humTD03pl	0.100	6.8%	5.6%	5.2%	0.247	0.118	9.0%	50
humTD03l	0.051	4.8%	4.4%	3.1%	0.152	0.077	3.6%	50

Table 12: Impact of Topic Distillation Techniques on Precision@10

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
u (upl - pl)	0.072	(0.035, 0.113)	25-5-20	0.600 (7), 0.400 (32)
p (pl - l)	0.012	(−0.009, 0.035)	8-6-36	0.300 (48), 0.200 (15)
l (upl - up)	0.004	(−0.019, 0.027)	7-4-39	0.300 (43), −0.300 (31)
h (uhpl - upl)	−0.026	(−0.049,−0.003)	5-13-32	−0.300 (7), −0.200 (9)

matching (perhaps the topics this year happened to be part of the title of the desired page more often than last year).

Table 10 isolates the components of the 'u' factor. 'r' denotes the weight assigned to the URL_TYPE values (ROOT, SUBROOT, PATH) and 'd' denotes the weight assigned to the URL_DEPTH values ('u' was 'r10d5' and is in Table 8). A small weight on either the url depth or type increased the home page score substantially without a significant drop in the named page score (as per the 'd5' and 'r5' rows). So it may be reasonable to include a small weight on url structure in a general web page search system, regardless of the expected frequency ratio of home page and named page queries. Higher weights may be reasonable if home page queries are expected to be a lot more common.

Table 11 shows the scores of the submitted topic distillation runs in descending order by Precision@10. The humTD03upl run had the highest Precision@10 score of any submitted run from the 23 groups, even though its score means it found on average just more than 1 key resource page in the first 10 rows (the judgements contained 8 key resource pages per topic on average). The topics were broad (e.g. "science" was an example in the task guidelines) and the top retrieved rows may have been filled with many more pages that were "relevant" to the topic even though they were not judged "key resources" by the assessors.

Tables 12, 13 and 14 show the impact of the submitted topic distillation techniques on Precision@10, average precision and Precision@R respectively:

- The 'u' factor (extra weight for URL type and depth) increased Precision@10 by 7 points and produced a statistically significant increase for all 3 examined measures. This is not surprising because the key resources were required to be home pages this year.
- The 'p' factor (extra weight for HTML properties and phrases in properties) did not have a significant impact on Precision@10, but Tables 13 and 14 show it led to a significant increase in the average

Table 13: Impact of Topic Distillation Techniques on Average Precision

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
p (pl - l)	0.049	(0.002, 0.111)	39-9-2	0.956 (24), 0.944 (17)
u (upl - pl)	0.038	(0.009, 0.069)	35-12-3	0.343 (49), 0.337 (18)
l (upl - up)	0.019	(−0.011, 0.066)	16-22-12	1.000 (17), −0.152 (15)
h (uhpl - upl)	−0.041	(−0.079,−0.012)	13-35-2	−0.750 (17), −0.187 (7)

Table 14: Impact of Topic Distillation Techniques on R-Precision

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
u (upl - pl)	0.059	(0.021, 0.100)	20-3-27	0.667 (49), 0.462 (7)
p (pl - l)	0.054	(0.004, 0.119)	11-3-36	1.000 (24), 1.000 (17)
l (upl - up)	0.016	(−0.018, 0.065)	6-3-41	1.000 (17), −0.250 (15)
h (uhpl - upl)	−0.042	(−0.095,−0.001)	7-15-28	−1.000 (17), −0.333 (49)

precision and Precision@R measures.

- The 'l' factor (linguistic expansion (inflections) from lexical English stemming) made little difference for most topics except for some measures for topic 17 ("Polygraphs").
- The 'h' factor (even more extra weight for URL type) had a significant negative impact on the examined measures.

Overall, the impacts on the distillation scores were much more like the impacts on the home page finding scores than the named page finding scores.

4 Genomic Retrieval

For the primary task of the Genomics Track (find all records focusing on the named gene), the MEDLINE data consisted of 525,938 documents (records), all in one file ("trec-medline") of 1,158,771,473 bytes (1.1 GB) uncompressed. The average record length was 2203 bytes. More information should be in the track overview paper.

4.1 Indexing

The cTREC text reader (described in the Robust section) was enhanced to include a /M option for identifying the MEDLINE records (documents) during table expansion from the "trec-medline" file. For the individual records, the /p option of cTREC was used, i.e. we just passed through all of the text for indexing, including identifiers such as "UI", "PMID", "MH" etc. (Maybe next year we will enhance the text reader to populate columns from particular fields, such as the Title, allowing experiments with the record structure like we did for the web data.)

A different stopfile, mynum.stp, was used for this task. It contained just one instruction, AL = "0-9", which means to treat the digits 0 to 9 as alphabet characters. For example, this would cause the symbol "CDKN1A" to be indexed as 1 term instead of 3. Experiments on the training queries found the scores were a little higher with this indexing change. The mynum.stp stopfile did not contain any stop words as the training queries did not seem to use much natural language.

Punctuation characters, including hyphens and parentheses, were still treated as term separators.

4.2 Searching

The submitted runs used IS_ABOUT queries based on combining just 5 of the 8 query fields: the 2 name fields (OFFICIAL_GENE_NAME, PREFERRED_GENE_NAME) and the 3 symbol fields (OFFICIAL_SYMBOL, ALIAS_SYMBOL, PREFERRED_SYMBOL). The other 3 fields were omitted (PREFERRED_PRODUCT, ALIAS_PROT, PRODUCT) because they were found to be harmful on the training topics. Also, the species information was ignored for the submitted runs.

The submitted humG03ns run gave equal weight to the five fields. An example SearchSQL query is below (from run humG03ns test topic 1). The parentheses between query fields were just added for readability and did not affect the IS_ABOUT search:

Table 15: Precision of Genomics Runs

Run	AvgP	P@5	P@10	P@20	Rec0	Rec30	P@R	Topics
humG03ns	0.175	16.4%	14.8%	11.7%	0.394	0.239	15.3%	50
humG03ns5	0.185	18.0%	15.8%	12.3%	0.399	0.257	16.7%	50
base (diag.)	0.312	27.2%	23.6%	18.0%	0.599	0.420	28.9%	50
+5x phrases	0.356	33.2%	24.8%	20.0%	0.613	0.463	31.6%	50

```
SELECT RELEVANCE('V2:3') AS REL, DOCNO
FROM med03n
WHERE FT_TEXT IS_ABOUT 'activating transcription factor 2 ()
  ATF2 () HB16 () CREB2 () TREB7 () CRE-BP1'
ORDER BY REL DESC
```

The submitted humG03ns5 run gave 5 times the weight to the three symbol fields (by repeating them 5 times in the query, rather than using the WEIGHT clause), which was modestly helpful on the training topics (though not significantly so).

Inflections from stemming were disabled for both runs. The document length importance was set to 0 for both runs. More details of the relevance ranking are in the Robust and Web sections.

4.3 Results

Table 15 shows various scores of the submitted runs, humG03ns and humG03ns5 (the column headings are explained in the Robust section).

At the conference, some groups found that filtering by species (organism) was helpful, presumably because of the artificial way the relevance assessments were created for this first Genomics task. The NLM paper [6] described how to convert the species name in the topic statement to the MH field in the MEDLINE records (map 'Homo sapiens' to 'Human', map 'Mus musculus' to 'Mice', map 'Rattus norvegicus' to 'Rats', map 'Drosophila melanogaster' to 'Drosophila'). The NRC reported that just 10 of the official right answers were discarded if restricting to fields of the same organism [2].

The diagnostic "base run" is the same as humG03ns except that it adds a phrase-match restriction to just include documents of the specified species, e.g. "AND (FT_TEXT CONTAINS 'MH - Human' WEIGHT 0)" was added to the query if the species was 'Homo sapiens'. It was assigned "WEIGHT 0" so that it would not affect the relevance calculation. Table 15 shows that the base run scored a 0.312 mean average precision, an increase of more than 13 points over humG03ns.

The "+5x phrases" diagnostic run of Table 15 additionally boosted the scores of records which contained any of the query fields as complete phrases, by use of the CONTAINS predicate. In the CONTAINS predicate, hyphenated terms match not just terms separated with different punctuation or white space, but also concatenations of the terms (e.g. a CONTAINS search for 'CRE-BP1' would additionally match not just 'CRE(BP1)', 'CRE BP1', etc., but also 'CREBP1'). The WHERE clause for topic 1 was

```
WHERE ( FT_TEXT IS_ABOUT 'activating transcription factor 2 ()
              ATF2 () HB16 () CREB2 () TREB7 () CRE-BP1'
   OR (FT_TEXT CONTAINS 'activating transcription factor 2' WEIGHT 5)
   OR (FT_TEXT CONTAINS 'ATF2' WEIGHT 5)
   OR (FT_TEXT CONTAINS 'HB16' WEIGHT 5)
   OR (FT_TEXT CONTAINS 'CREB2' WEIGHT 5)
   OR (FT_TEXT CONTAINS 'TREB7' WEIGHT 5)
   OR (FT_TEXT CONTAINS 'CRE-BP1' WEIGHT 5) )
  AND (FT_TEXT CONTAINS 'MH - Human' WEIGHT 0)
```

Table 16 compares a number of diagnostic runs to the base run (always subtracting the base run's scores in average precision from the listed run). For example, the first row shows that the "+5x phrases" run

Table 16: Impact of Genomics Techniques on Average Precision

Experiment	AvgDiff	95% Confidence	vs.	2 Largest Diffs (Topic)
+5x phrases	0.044	(0.015, 0.076)	28-18-4	0.441 (24), 0.266 (23)
+2x phrases	0.043	(0.016, 0.073)	29-17-4	0.441 (24), 0.316 (23)
+1x phrases	0.037	(0.011, 0.066)	32-14-4	0.441 (24), 0.314 (23)
2x sym	0.027	(0.007, 0.052)	28-19-3	0.441 (24), 0.205 (4)
idf squared (V2:4)	0.020	(−0.006, 0.048)	24-23-3	0.441 (24), −0.231 (18)
5x sym	0.019	(−0.010, 0.051)	22-23-5	0.441 (24), 0.280 (15)
phrases only	0.002	(−0.058, 0.058)	28-19-3	−0.733 (7), −0.631 (20)
DLEN 500	−0.006	(−0.014, 0.001)	17-29-4	−0.119 (16), −0.074 (18)
stemming on	−0.012	(−0.036, 0.005)	11-27-12	−0.463 (27), −0.167 (16)
omit names	−0.030	(−0.079, 0.016)	16-32-2	−0.733 (7), 0.441 (24)
number parsing	−0.038	(−0.071,−0.004)	16-30-4	0.334 (11), −0.331 (9)
all fields	−0.055	(−0.083,−0.031)	9-35-6	−0.382 (31), −0.312 (36)
vector species	−0.069	(−0.102,−0.034)	7-41-2	−0.333 (27), −0.329 (28)
omit symbols	−0.113	(−0.152,−0.075)	8-40-2	−0.489 (31), −0.429 (19)
terms count (2:2)	−0.136	(−0.182,−0.094)	5-44-1	−0.717 (27), −0.507 (29)
omit species	−0.137	(−0.177,−0.099)	2-46-2	−0.489 (20), −0.489 (27)
hits count (2:1)	−0.199	(−0.246,−0.154)	3-47-0	−0.619 (27), −0.543 (47)

scored on average 4 points higher than the base run (0.356 minus 0.312 is 0.044), and this difference was statisticially significant (see the Robust section for a detailed explanation of the column headings of Table 16).

Phrasing: The "+2x phrases" and "+1x phrases" runs used 'WEIGHT 2' and 'WEIGHT 1' for the phrases instead of 'WEIGHT 5', and they also produced significant 4 point gains. The "phrases only" run just used the phrases (dropping the IS_ABOUT predicate) and scored about the same as the base run on average, though with a lot of variance. The "number parsing" run used a table with the default parsing of alphanumerics (e.g. "CDKN1A" would be treated as 3 terms (CDKN, 1, A) instead of 1), in a sense removing the natural phrasing of symbols, and the 4 point drop in score passed the significance test. (Note that if the symbols matched as phrases when names did not, the effect would be the same as just increasing the symbol weight (described below), which may be why topic 24 shows a similar increase in Table 16 for both phrases and symbol weighting.) There is probably room for improvement in this term matching area (e.g. a search for 'CDKN1A' will not match 'CDKN 1A' with the parsing rules used for this task).

Query fields: The "2x sym" and "5x sym" runs were the same as the diagnostic base run except that the symbols were each listed twice and five times (respectively) to boost their impact on the score. Table 16 shows the 3 point gain of "2x sym" was statistically significant, while the 2 point gain of "5x sym" was not. Just using the symbols (omitting the name fields) scored 3 points lower on average (as per the "omit names" run), but with a lot of variance. Just using the names and not the symbols scored a significant 11 points lower (as per the "omit symbols" run). Adding in the 3 other fields (preferred product, product, alias prot) scored a significant 5 points lower (as per the "all fields" run). Overall, it appears the symbols are the most useful of the query fields for this task, though perhaps we're not making as effective use of the names as we could (as the phrase experiments suggested).

Relevance ranking: Squaring the importance of inverse document frequency to the relevance calculation (by using SearchServer relevance method 'V2:4' instead of 'V2:3') scored 2 points higher, but did not quite pass the significance test, as per the listed "idf squared (V2:4)" run. Enabling document length normalization or matching of inflections from English stemming made little difference for this task as per the listed "DLEN 500" and "stemming on" runs. Simpler ranking techniques, such as just counting the number of query terms matched (relevance method '2:2') or simply counting all the matches in a record (relevance method '2:1') scored dramatically lower (14 and 20 points respectively, as per the listed "terms count (2:2)" and "hits count (2:1)" runs) indicating that a combination of term frequency dampening and inverse document frequency is

still valuable for this task (though we have not separated the impact of these techniques).

As previously mentioned, not restricting to the species given in the topic scored more than 13 points lower as per the listed "omit species" run (the difference of the humG03ns run and the base run). Adding the species to the IS_ABOUT vector instead of using a strict CONTAINS match scored 7 points lower (as per the listed "vector species" run). At the conference it was stated that in a real task it can be useful to find the gene in different species, so these species results apparently are examples of misleading conclusions from the artificial nature of the judgements used this year.

References

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

[2] Berry de Bruijn and Joel Martin. Finding Gene Function using LitMiner. Institute for Information Technology, National Research Council of Canada. Notebook paper in draft TREC 2003 Conference Proceedings.

[3] Bradley Efron and Robert J. Tibshirani. An Introduction to the Bootstrap. 1993. Chapman & Hall/CRC.

[4] The .GOV Test Collection. http://www.ted.cmis.csiro.au/TRECWeb/govinfo.html

[5] Andrew Hodgson. Converting the Fulcrum Search Engine to Unicode. In Sixteenth International Unicode Conference, Amsterdam, The Netherlands, March 2000.

[6] Mehmet Kayaalp et al. Methods for accurate retieval of MEDLINE citations in functional genomics. National Library of Medicine. Notebook paper in draft TREC 2003 Conference Proceedings.

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

[8] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford. (City University.) Okapi at TREC-3. In D. K. Harman, editor, Overview of the Third Text REtrieval Conference (TREC-3). NIST Special Publication 500-226. http://trec.nist.gov/pubs/trec3/t3_proceedings.html

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

[10] Stephen Tomlinson and Tom Blackwell. Hummingbird's Fulcrum SearchServer at TREC-9. In E. M. Voorhees and D. K. Harman, editors, Proceedings of the Ninth Text REtrieval Conference (TREC-9). NIST Special Publication 500-249. http://trec.nist.gov/pubs/trec9/t9_proceedings.html

[11] Stephen Tomlinson. Experiments in 8 European Languages with Hummingbird SearchServer™ at CLEF 2002. In Carol Peters, editor, Working Notes for the CLEF 2002 Workshop. http://clef.iei.pi.cnr.it:2002/workshop2002/WN/26.pdf

[12] Stephen Tomlinson. Experiments in Named Page Finding and Arabic Retrieval with Hummingbird SearchServer™ at TREC 2002. In E. M. Voorhees and Lori P. Buckland, editors, Proceedings of the Eleventh Text REtrieval Conference (TREC 2002). NIST Special Publication 500-251. http://trec.nist.gov/pubs/trec11/t11_proceedings.html

[13] Stephen Tomlinson. Hummingbird SearchServer™ at TREC 2001. In E. M. Voorhees and D. K. Harman, editors, Proceedings of the Tenth Text REtrieval Conference (TREC 2001). NIST Special Publication 500-250. http://trec.nist.gov/pubs/trec10/t10_proceedings.html

[14] Stephen Tomlinson. Lexical and Algorithmic Stemming Compared for 9 European Languages with Hummingbird SearchServer™ at CLEF 2003. In Carol Peters, editor, Working Notes for the CLEF 2003 Workshop. http://clef.iei.pi.cnr.it/2003/WN_web/19.pdf

[15] Thijs Westerveld, Wessel Kraaij and Djoerd Hiemstra. Retrieving Web Pages using Content, Links, URLs and Anchors. In E. M. Voorhees and D. K. Harman, editors, Proceedings of the Tenth Text REtrieval Conference (TREC 2001). NIST Special Publication 500-250. http://trec.nist.gov/pubs/trec10/t10_proceedings.html