The TREC 2005 Terabyte Track

Charles L. A. Clarke
University of Waterloo
claclark@plg.uwaterloo.ca

Falk Scholer
RMIT
fscholer@cs.rmit.edu.au

Ian Soboroff
NIST
ian.soboroff@nist.gov

1 Introduction

The Terabyte Track explores how retrieval and evaluation techniques can scale to terabyte-sized collections, examining both efficiency and effectiveness issues. TREC 2005 is the second year for the track. The track was introduced as part of TREC 2004, with a single adhoc retrieval task. That year, 17 groups submitted 70 runs in total. This year, the track consisted of three experimental tasks: an adhoc retrieval task, an efficiency task and a named page finding task. 18 groups submitted runs to the adhoc retrieval task, 13 groups submitted runs to the efficiency task, and 13 groups submitted runs to the named page finding task. This report provides an overview of each task, summarizes the results and discusses directions for the future. Further background information on the development of the track can be found in last year's track report [4].

2 The Document Collection

All tasks in the track use a collection of Web data crawled from Web sites in the gov domain during early 2004. We believe this collection ("GOV2") contains a large proportion of the crawlable pages present in gov at that time, including HTML and text, along with the extracted contents of PDF, Word and postscript files. The collection is 426GB in size and contains 25 million documents. In 2005, the University of Glasgow took over the responsibility for distributing the collection. In 2004, the collection was distributed by CSIRO, Australia, who assisted in its creation.

3 The Adhoc Task

An adhoc task in TREC investigates the performance of systems that search a static set of documents using previously-unseen topics. For each topic, participants create a query and generate a ranked list of the top 10,000 documents for that topic. For the 2005 task, NIST created and assessed 50 new topics. An example is provided in figure 1.

As is the case for most TREC adhoc tasks, a topic describes the underlying information need in several forms. The title field essentially contains a keyword query, similar to a query that might be entered into a Web search engine. The description field provides a longer statement of the topic requirements, in the form of a complete sentence or question. The narrative, which may be a full

```
<top>
<num> Number: 756

<title> Volcanic Activity

<desc> Description:
Locations of volcanic activity which occurred within the present day
boundaries of the U.S. and its territories.

<narr> Narrative:
Relevant information would include when volcanic activity took place,
even millions of years ago, or, on the contrary, if it is a possible
future event.

</top>
```

Figure 1: *Adhoc Task Topic 756*

paragraph in length, supplements the other two fields and provides additional information required to specify the nature of a relevant document.

For the adhoc task, an experimental run consisted of the top 10,000 documents for each topic. To generate a run, participants could create queries automatically or manually from the topics. For most experimental runs, participants could use any or all of the topic fields when creating queries from the topic statements. However, each group submitting any automatic run was required to submit at least one automatic run that used only the title field of the topic statement. Manual runs were encouraged, since these runs often add relevant documents to the evaluation pool that are not found by automatic systems using current technology. Groups could submit up to four runs.

The pools used to create the relevance judgments were based on the top 100 documents from two adhoc runs per group, along with two efficiency runs per group. This yielded an average of 906 documents judged per topic (min 347, max 1876). Assessors used a three-way scale of "not relevant", "relevant", and "highly relevant". A document is considered relevant if any part of the document contains information which the assessor would include in a report on the topic. It is not sufficient for a document to contain a link that appears to point to a relevant web page, the document itself must contain the relevant information. It was left to the individual assessors to determine their own criteria for distinguishing between relevant and highly relevant documents. For the purpose of computing the effectiveness measures, which require binary relevance judgments, the relevant and highly relevant documents were combined into a single "relevant" set.

In addition to the top 10,000 documents for each run, we collected details about the hardware and software configuration, including performance measurements such as total query processing time. For total query processing time, groups were asked to report the time required to return the top 20 documents, not the time to return the top 10,000. It was acceptable to execute a system twice for each run, once to generate the top 10,000 documents and once to measure the execution time for the top 20 documents, provided that the top 20 documents were the same in both cases.

Figure 2 provides an summary of the results obtained by the eight groups achieving the best results according to the bpref effectiveness measure [3]. When possible, we list two runs per group:

Group	**Run**	**bpref**	**MAP**	**p@20**	**CPUs**	**Time (sec)**
umass.allan	indri05AdmfS	0.4279	0.3886	0.5980	6	5162
	indri05Aql	0.3714	0.3252	0.5650	6	62
hummingbird.tomlinson	humT05xle	0.4264	0.3655	0.6230	1	50000
	humT05l	0.3659	0.3154	0.5800	1	5700
uglasgow.ounis	uogTB05SQE	0.4178	0.3755	0.6180	8	1000
uwaterloo.clarke	uwmtEwtaPt	0.3884	0.3451	0.5760	2	63
	uwmtEwtaD02t	0.2887	0.2173	0.4490	2	3
umelbourne.anh	MU05TBa2	0.3771	0.3218	0.5730	1	10
ntu.chen	NTUAH2	0.3760	0.3233	0.5630	1	734
	NTUAH1	0.3555	0.3023	0.5400	1	270
dublincityu.gurrin	DCU05AWTF	0.3603	0.3021	0.5600	5	120
tsinghua.ma	THUtb05SQWP1	0.3553	0.3032	0.5330	1	1800

Figure 2: *Adhoc Results (top eight groups by bpref)*

the run with the highest bpref and the run with the fastest time. The first two columns of the table identify the group and run. The next three columns provide the values of three standard effective measures for each run: bpref, mean average precision (MAP) and precision at 20 documents (p@20) [3]. The last two columns list the number of CPUs used to generate the run and the total query processing time. When the fastest and best runs are compared within groups, the trade-off between efficiency and effectiveness is apparent. This trade-off is further explored in the discussion of the efficiency results.

4 Efficiency Task

The efficiency task extends the adhoc task, providing a vehicle for discussing and comparing efficiency and scalability issues in IR systems by defining better methodology to determine query processing times. Nonetheless, the validity of direct comparisons between groups is limited by the range of hardware used, which varies from desktop PCs to supercomputers. Thus, participants are encouraged to compare techniques within their own systems or to compare the performance of their systems to that of public domain systems.

Ten days before the new topics were released for the adhoc task, NIST released a set of 50,000 efficiency test topics, which were extracted from the query logs of an operational search engine. Figure 3 provides some examples. The title fields from the new adhoc topics were seeded into this topic set, but were not distinguished in any way. Participating groups were required to process these topics automatically; manual runs were not permitted for this task.

Participants executed the entire topic set, reporting the top-20 results for each query and the total query processing time for the full set. Query processing time included the time to read the topics and write the final submission file. The processing of topics was required to proceed sequentially, in the order the topics appeared in the topic file. To measure effectiveness, the results corresponding to the adhoc topics were extracted and added into the evaluation pool for the adhoc task. Since the efficiency runs returned only the top 20 documents per topic, they did not substantially increase the pool size.

```
7550:yahoo
7551:mendocino and venues
7552:creative launcher
7553:volcanic activity
7554:shorecrest
7555:lazy boy
7556:los bukis deseo download free
7557:online surveys
7558:wholesale concert tickets
```

Figure 3: *Efficiency Task Topics 7550 to 7558*

Group	**Run**	**p@20**	**CPUs**	**Time (sec)**	**US$ Cost**
uwaterloo.clarke	uwmtEwtePTP	0.5780	2	54701	3800
	uwmtEwteD10	0.3900	2	1371	3800
umelbourne.anh	MU05TBy1	0.5620	8	2145	6000
	MU05TBy3	0.5550	8	1201	6000
hummingbird.tomlinson	humTE05i4ld	0.5490	1	219354	5000
	humTE05i5	0.4460	1	39506	5000
umass.allan	indri05Eql	0.5490	1	71700	1500
	indri05EqlD	0.5490	6	24720	9000
rmit.scholer	zetdir	0.5410	1	11565	1200
	zetdist	0.5300	8	2901	6000
dublincityu.gurrin	DCU05DISTWTF	0.5290	5	48375	13125
	DCU05WTFQ	0.4660	1	17730	2625
ntu.chen	NTUET2	0.5180	1	186900	2400
	NTUET1	0.5150	1	183200	2400
upisa.attardi	pisaEff2	0.4350	23	12898	10000
	pisaEff4	0.3420	23	7158	10000

Figure 4: *Efficiency Results (top eight groups by p@20)*

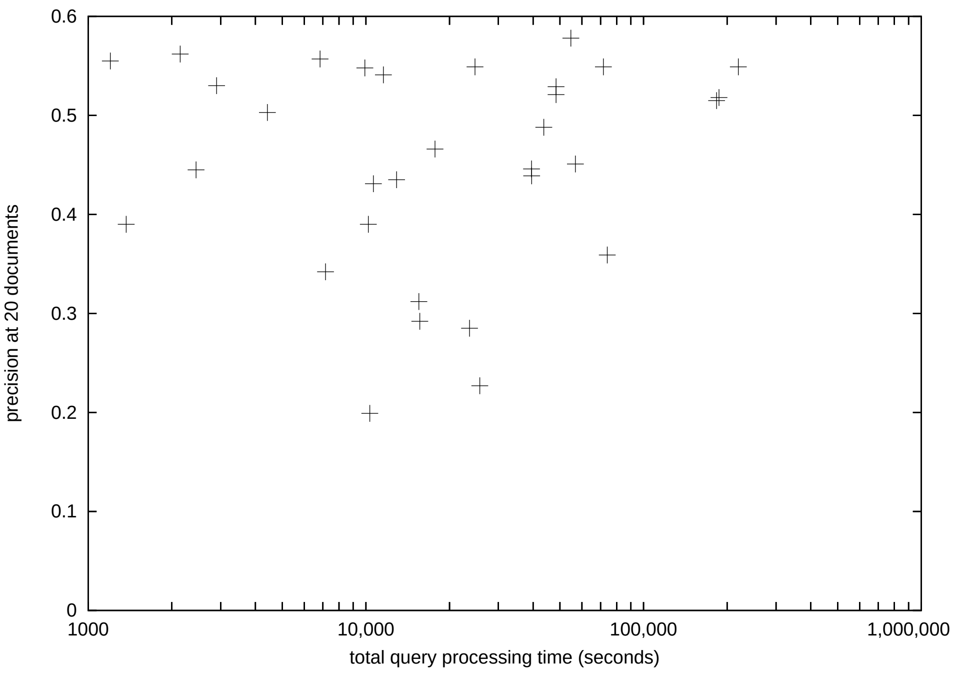

Figure 5: *Efficiency vs. Effectiveness*

Figure 4 summarizes the results for the eight groups achieving the best results, based on p@20. Once again, the figure lists both the best and fastest run from each group. In addition to the query processing times and number of CPUs, the table also includes the estimate of total hardware cost provided by each participating group.

To illustrate the range of results seen within the track, and to provide a sense of the trade-offs between efficiency and effectiveness, figure 5 plots p@20 against total query processing time for all 32 runs submitted to the efficiency track. Note that a log scale is used to plot query processing times. The range in both dimensions is quite dramatic.

The results plotted in figure 5 were generated on a variety of hardware platforms, with different costs and configurations. To adjust for these differences, we attempted two crude normalizations. Figure 6 plots p@20 against total query processing time, normalized by the number of CPUs. The normalization was achieved simply by multiplying the time by the number of CPUs. Figure 7 plots p@20 against total query processing time, normalized by hardware cost, with the times adjusted to a typical uniprocessor server machine costing $2,000. In this case, the normalization consisted of multiplying the time by the cost and dividing by 2,000. Both normalizations have the effect of moving the points sharply away from the upper left-hand corner, making the trade-offs in this area more apparent.

5 Named Page Finding Task

Named page finding is a navigational search task, where a user is looking for a particular resource. In comparison to an adhoc task, a topic for a named page finding task usually has one answer: the resource specified in the query. The objective of the task, therefore, is to find a particular page in

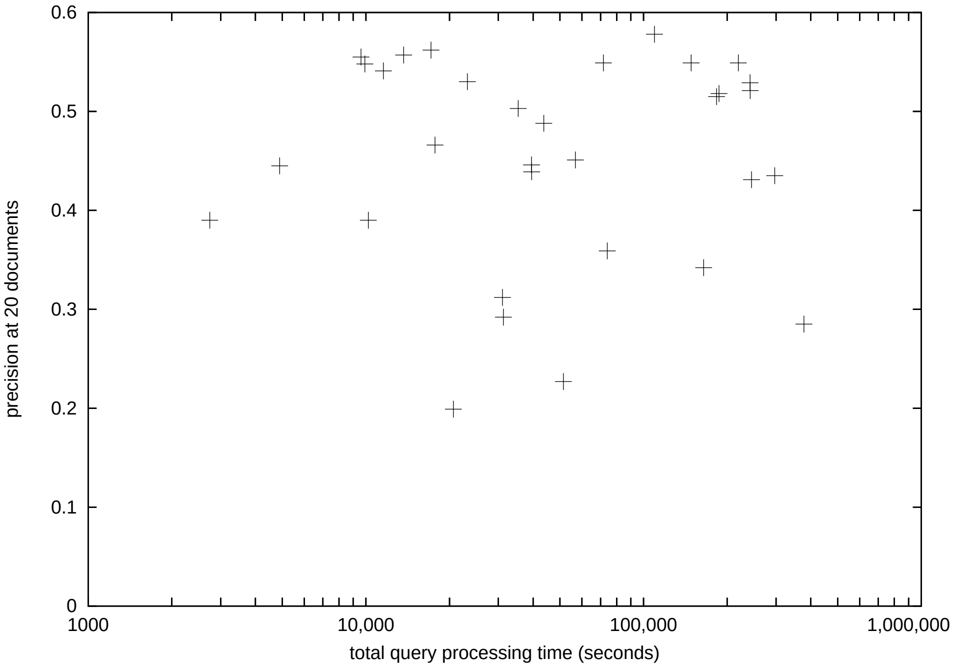

Figure 6: *Efficiency vs. Effectiveness (normalized by number of CPUs)*

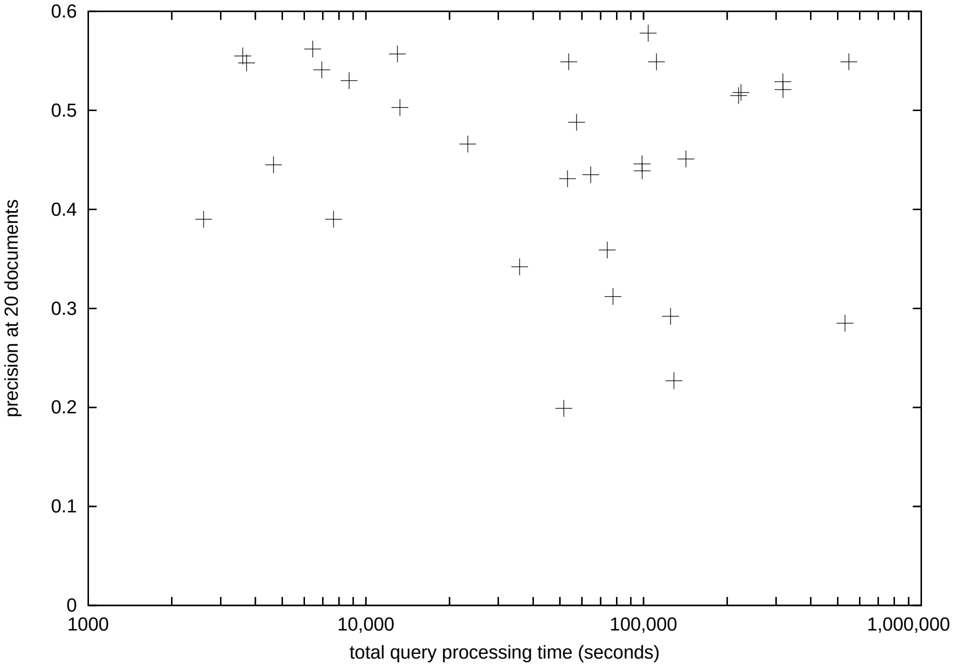

Figure 7: *Efficiency vs. Effectiveness (normalized by hardware cost)*

the GOV2 collection, given a topic that describes it. For example, the query "`fmcsa technical support`" would be satisfied by the Federal Motor Carrier Administration's technical support page.

Named page topics were created by envisaging a bookmark recovery process. Participants were presented with a randomly selected page from the GOV2 collection. If the page had any identifiable features, and looked like something that a real user might want to read, remember, and re-find at a later point in time, the participant was asked to write a query. Specifically, the task was: "write a query to retrieve the page, as if you had seen it 15 minutes ago, and then lost it". This process resulted in an initial list of 272 queries, each with one corresponding target page. After manual inspection, 20 of these were discarded because they were "topic finding" in nature. We believe that collection browsing facilities would have made topic creation far easier than merely viewing successive random pages, but this facility was not available at the time of topic creation.

Although named page topics are typically assumed to have a single correct answer, the topic creation process outlined above raises two issues: first, there may be exact duplicates of the target document in the collection; and second, the target may not be specified clearly enough to rule out topically similar pages as plausible answers.

The first issue was resolved by searching for near-exact duplicates of the target pages within the GOV2 collection. This was done with the DECO system [1], which uses a lossless fingerprinting technique for the detection of duplicate documents. Pools formed from the top 1000 answers from all 42 runs that were submitted for this year's task (around 1.5 million unique documents) were searched. All near-exact duplicates were included in the qrels file.

In the context of past TREC Web Tracks, the second issue was sometimes resolved by requiring the creation of "omniscient" queries; that is, each query is tested and, if it retrieves similar (but not identical) documents in the collection that could also be considered to be plausible answers, it is discarded. However, discarding such queries distances the experimental process from a real-world web search task: a user generally does not know in advance if a named page query is specific enough to only identify a single resource. For the Terabyte Track, we therefore chose to retain such queries, and treated them as having a single "correct" answer. As a result of the change in methodology, we expected that the named page finding task would be harder than previously experienced.

Of the 252 topics used for the named page finding task, 187 have a single relevant answer (that is, there are no exact duplicates that match the canonical named page in the answer pool). However, some pages repeat often in the collection (the highest number of duplicate answer pages were identified for topic 778, with 4525 repeats).

Figure 8 summaries the results of the named page finding task. The performance of the runs is evaluated using three metrics:

- **MRR**: The mean reciprocal rank of the first correct answer.
- **% Top 10**: The proportion of queries for which a correct answer was found in the first 10 search results.
- **% Not Found**: The proportion of queries for which no correct answer was found in the results list.

The figure lists the best run from the top eight groups by MRR. In addition, the figure indicates the runs that exploit link analysis techniques (such as pagerank), anchor text, and document structure (such as giving greater weight to terms appearing in titles). While reasonable results can be achieved without exploiting these web-related document characteristics, most of the top runs incorporate one or more of them.

Group	Run	MRR	% Top 10	% Not Found	CPUs	Time (sec)	Links?	Anchors?	Structure?
tsinghua.ma	THUtb05npW15	0.463	61.5	17.9	1	5400	N	Y	N
umass.allan	indri05Nmpsd	0.441	58.3	17.1	6	16200	Y	Y	Y
uglasgow.ounis	uogNP05baseN	0.401	54.8	19.8	64	3840	N	Y	Y
ntu.chen	NTUNF3	0.388	51.2	19.4	1	46200	N	Y	Y
hummingbird.tomlinson	humTN05pl	0.378	50.0	19.8	1	14000	N	N	Y
uwaterloo.clarke	uwmtEwtnpP	0.366	50.8	20.6	2	944	N	N	N
uamsterdam.mueller	UAmsT05n3SUM	0.365	48.8	22.6	2	13239	N	Y	N
yorku.huang	york05tNa1	0.329	44.4	25.4	1	10365	N	N	N

Figure 8: *Named Page Finding Results (top 8 groups by MRR)*

6 The Limits of Pooling

Aside from scaling TREC and research retrieval systems, a primary goal of the terabyte track is to determine if the Cranfield paradigm of evaluation scales to very large collections, and to propose alternatives if it doesn't. In the discussions which led to the current terabyte track, the hypothesis was that in very large collections we would not be able to find a good sample of relevant documents using the pooling method. If judgments are not sufficiently complete, then runs which were not pooled will retrieve unjudged documents, making those runs difficult to measure. Depending on the run and the reason that the judgments are incomplete, this can result in a biased test collection. According to MAP, unjudged documents are assumed irrelevant and a run may score lower than it should. The bpref measure can give artificially high scores to a run if they do not retrieve sufficient judged irrelevant documents.

One method for determining if pooling results in insufficiently complete judgments is to remove a group's pooled runs from the pools, and measure their runs as if they had not contributed their unique relevant documents. This test does reveal that the terabyte collections should be used with some caution. For 2004, the mean difference in scores across runs when the run's group is held out is 9.6% in MAP, and the maximum is 45.5%; on the other hand, this was the first year of the track, and overall system effectiveness was not very good. This year, the mean difference is 3.9% and the maximum is 17.7%, much more reasonable but still somewhat higher than we see in newswire.

Another approach is to examine the documents to see if they seem to be biased towards any particular retrieval approach. We have found that the relevant documents in both terabyte collections have a very high occurrence of the title words from the topics, and that this occurrence is much higher than we see in news collections for ad hoc retrieval. More formally, the *titlestat* measure for a set of documents D is defined to be the percentage of documents in D that contain a title word, computed as follows. For each word in the title of a topic that is not a stop word, calculate the percentage of the set D that contain the word, normalized by the maximum possible

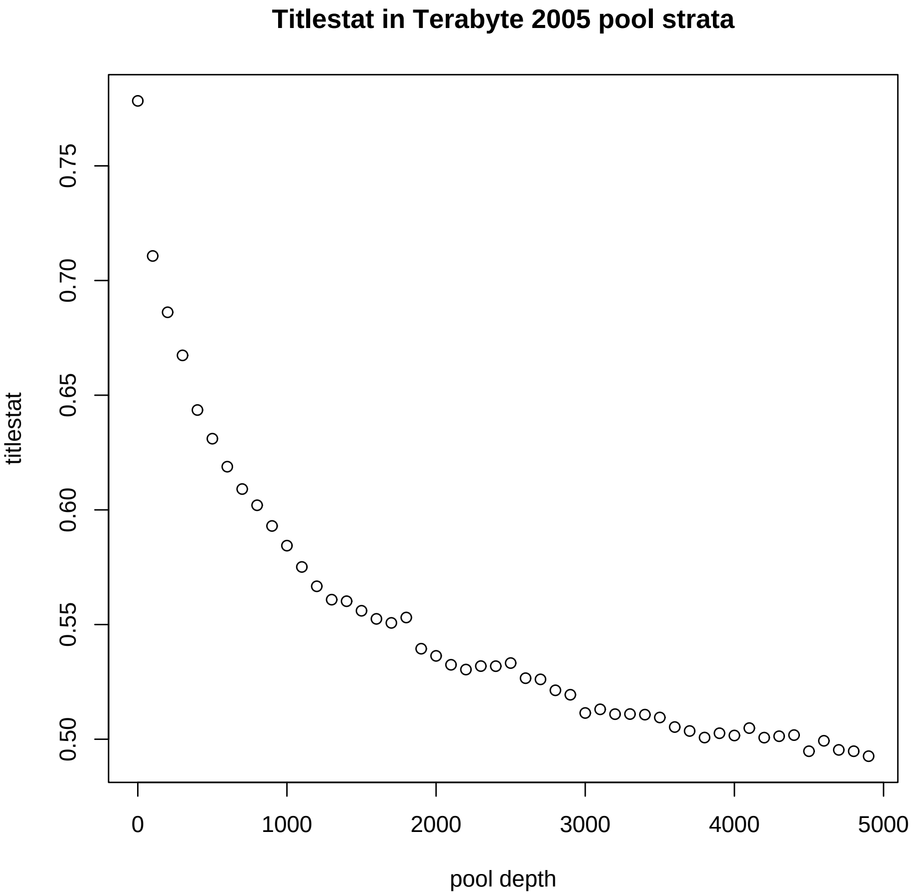

Figure 9: Titlestat in 100-document strata of the 2005 terabyte pools.

percentage.[1] For a topic, this is averaged over all words in the title, and for a collection, averaged over all topics. A maximum of 1.0 occurs when all documents in D contain all topic title words; 0.0 means that no documents contain a title word at all. Titlestat can be thought of as the occurrence of the average title word in the document set.

Titlestat can be measured for any set of documents. For the relevant documents (*titlestat_rel*)in the terabyte collections, we obtain 0.889 for the 2004 collection and 0.898 for 2005. In contrast, the TREC-8 ad hoc collection (TREC CDs 4 and 5 less the Congressional Register) has a *titlestat_rel* of 0.688. For the WT10g web collection, the TREC-9 ad hoc task relevant documents have a *titlestat_rel* of 0.795, and TREC-10 is 0.761.

Why are the terabyte titlestats so high? We feel that this is directly due to the size of the collection. In nearly any TREC collection, the top ranked documents are going to reflect the title words, since (a) title-only runs are often required, (b) even if they are not required, the title words are often used as part of any query, and (c) most query expansion will still weight the original query

[1] In rare cases, a title word will have a collection frequency smaller than $|D|$.

terms (i.e., the title words) highly. So it's certainly expected that the top ranks will be dominated by documents that contain the title words. For GOV2, the collection is so large that the title words have enormous collection frequency compared to the depth of the assessment pool. The result of this is that the pools are completely filled with title-word documents, and documents without title words are simply not judged.

Figure 9 illustrates this phenomenon using the titlestat of the pools, rather than of the judged relevant documents. The first point at $x = 0$ is the titlestat of the pool from depth 1-100, the pool depth used this year (0.778). In contrast, the titlestat of the TREC-8 pools is 0.429. Subsequent points in the graph show the titlestat of the pool from depth 101-200, 201-300, and so forth. Each pool is cumulative with respect to duplicates, meaning that if a document was pooled at a shallower depth it is not included in a deeper pool stratum. In order to get to lower titlestat depths, we would have had to pool very deep indeed. In any event, these titlestats indicate that the pools are heavily biased towards documents containing the title words, and may not fairly measure runs which do not use the title words in their query.

7 The Future of the Terabyte Track

Our analysis of title-word occurrence within the terabyte pools and relevance judgments indicates that the terabyte collections may be biased towards title-only runs. This is a serious concern for a TREC collection, and for the 2006 adhoc task we intend to pursue several strategies to build a more reusable test collection. In part, greater emphasis will be placed on the submission of manual runs, expanding the variety of relevant documents in the pools to include more documents that contain few or none of the query terms and increasing the re-usability of the collection. Users of the 2004 and 2005 collections should be very cautious. We recommend the use of multiple effectiveness measures (such as MAP and bpref) and careful attention to the number of retrieved unjudged documents.

In addition, the evaluation procedure may be modified to reduce the influence of *content-equivalent* documents in the collection. Using the 2004 topics as a case study, Bernstein and Zobel [2] present methods for identifying these near-duplicate documents and discover a surprisingly high level of inconsistency in their judging. Moreover, these near duplicates represent up to 45% of the relevant documents for given topics. This inconsistency and redundancy has a substantial impact on effectiveness measures, which we intend to address in the definition of the 2006 task.

Along with the adhoc task, we plan to run a second year of the efficiency and named page finding tasks, allowing groups to refine and test methods developed this year. In the case of the efficiency task, we are developing a detailed query execution procedure, with the hope of allowing more meaningful comparisons between systems.

Planning for 2006 is an ongoing process. As our planning progresses, it is possible that we may add an efficiency aspect to the named page finding task, and a "snippet retrieval" aspect to the adhoc retrieval task. A substantial expansion of the test collection remains a long-term goal, if the track continues in the future.

Acknowledgments

We thank Nick Craswell, David Hawking and others at CSIRO for their assistance with the creation of the GOV2 collection, and Doug Oard who hosted our Web crawler at the University of Maryland. The members of the Search Engine Group at RMIT University helped in the creation of the named page topics for this year's Terabyte Track. Finally, we thank Yaniv Bernstein, who kindly made his DECO software available for the identification of duplicate answer pages.

References

[1] Yaniv Bernstein and Justin Zobel. A scalable system for identifying co-derivative documents. In *Proceedings of the Symposium on String Processing and Information Retrieval*, pages 55–67, Padova, Italy, 2004.

[2] Yaniv Bernstein and Justin Zobel. Redundant documents and search effectiveness. In *Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management*, pages 736–743, Bremen, Germany, 2005.

[3] Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–32, Sheffield, UK, 2004.

[4] C. L. A. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2004 Terabyte Track. In *Proceedings of the Thirteenth Text REtrieval Conference*, Gaithersburg, MD, November 2004. NIST Special Publication 500-261. See `trec.nist.gov`.