University of Amsterdam and University of Twente at the TREC 2007 Million Query Track

Djoerd Hiemstra[1] **Jaap Kamps**[2,3] **Rianne Kaptein**[3] **Rongmei Li**[1]

[1] Database Group, University of Twente
[2] ISLA, Informatics Institute, University of Amsterdam
[3] Archives and Information Studies, Faculty of Humanities, University of Amsterdam

Abstract: In this paper, we document our submissions to the TREC 2007 Million Query track. Our main aim is to compare results of the earlier Terabyte tracks to the Million Query track. We submitted a number of runs using different document representations (such as full-text, title-fields, or incoming anchor-texts) to increase pool diversity. The initial results show broad agreement in system rankings over various measures on topic sets judged at both Terabyte and Million Query tracks, with runs using the full-text index giving superior results on all measures, but also some noteworthy upsets.

1 Introduction

The University of Amsterdam, in collaboration with the University of Twente, participated with the main aim to compare results of the earlier Terabyte tracks to the Million Query track. Specifically, what is the impact of shallow pooling methods on the (apparent) effectiveness of retrieval techniques? And what is the impact of substantially larger numbers of topics?

The rest of this paper is organized as follows. In Section 2, we detail the experimental set-up for the two tasks in the Terabyte track. In Section 3, we discuss our official submissions and results. Finally, we summarize our findings in Section 4.

2 Experimental Set-up

2.1 Retrieval set-up

Our retrieval system is based on the Lucene engine with a number of home-grown extensions [1, 6].

Indexes The Million Query track uses the GOV2 test collection, containing 25,205,178 documents (426 Gb uncompressed). The indexing approach is similar to our earlier experiments in the TREC Web and Terabyte tracks [2, 3, 4, 5]. We created three separate indexes for

Full-text the full textual content of the documents (covering the whole collection);

Titles the text in the title tags of each document, if present (covering 86% of the collection);

Anchors another anchor-texts index in which we unfold all relative links (covering 49% of the collection).

For the anchor text index, we normalized the URLs, and did not index repeated occurrences of the same anchor-text. As to tokenization, we removed HTML-tags, punctuation marks, applied case-folding, and mapped marked characters into the unmarked tokens. We used the Snowball stemming algorithm [7]. The main full document text index was created as a single, non-distributed index. The size of our full-text index is 61 Gb. Building the full-text index (including all further processing) took a massive 15 days, 6 hours, and 21 minutes.

Retrieval model For our ranking, we use either a vector-space retrieval model or a language model. Our vector space model is the default similarity measure in Lucene [6], i.e., for a collection D, document d and query q:

$$sim(q,d) = \sum_{t \in q} \frac{tf_{t,q} \cdot idf_t}{norm_q} \cdot \frac{tf_{t,d} \cdot idf_t}{norm_d} \cdot coord_{q,d} \cdot weight_t,$$

where $tf_{t,X} = \sqrt{\text{freq}(t,X)}$; $idf_t = 1 + \log \frac{|D|}{\text{freq}(t,D)}$; $norm_q = \sqrt{\sum_{t \in q} tf_{t,q} \cdot idf_t{}^2}$; $norm_d = \sqrt{|d|}$; and $coord_{q,d} = \frac{|q \cap d|}{|q|}$. Our language model is an extension to Lucene [1], i.e., for a collection D, document d and query q:

$$P(d|q) = P(d) \cdot \prod_{t \in q} \left((1-\lambda) \cdot P(t|D) + \lambda \cdot P(t|d) \right),$$

where $P(t|d) = \frac{tf_{t,d}}{|d|}$, $P(t|D) = \frac{\text{doc\_freq}(t,D)}{\sum_{t' \in D} \text{doc\_freq}(t',D)}$, and $P(d) = \frac{|d|}{\sum_{d' \in D} |d'|}$. The standard value for the smoothing parameter λ is 0.15. In previous years of the TREC Terabyte track, we found out that the GOV2 collection requires substantially less smoothing [2, 3]. That is, we use a value of λ close of 0.9 throughout.

Table 1: Statistics over judged and relevant documents per topic for million query track (top) and terabyte tracks (bottom).

	nr. of	per topic				
	topics	min	max	median	mean	st.dev
judged	1,692	6	147	40	41.15	6.68
relevant	1,455	1	52	10	12.36	9.69
highly rel.	710	1	44	3	5.44	6.28
judged	149	317	1,876	870	908.40	342.44
relevant	149	4	617	130	180.65	149.16
highly rel.	125	1	331	14	34.81	51.95

3 Experiments

3.1 Official runs

We submitted five runs before, and three runs after the official deadline. Two further runs were used to construct the official submissions. Only the five official submissions have been part of the pooling process.

We submitted two runs on the full-text index run, using the vector space model (`UAmsT07MTeVS`) and using the language model (`UAmsT07MTeLM`, not pooled).

Next, we submitted a plain title index run (`UAmsT07MTiLM`) and a plain anchor-text index run (`UAmsT07MAnLM`) both using the language model. We also have the similar runs using vector-space model, using the title index (`UAmsT07MTiVS`, not submitted) and the anchor-text index (`UAmsT07MAnVS`, not submitted).

These separate indexes can provide additional retrieval cues, for example, the anchor-texts provide a document representation completely disjoint from the document's text. Hence, we also submitted four run that combines different sources of evidence. First, a weighted CombSUM with relative weights of 0.6 (text), 0.2 (anchors), and 0.2 (titles) using the vector space model (`UAmsT07MSum6`) and the language model (`UAmsT07MSm6L`, not pooled). Second, a similar combination with relative weights of 0.8 (text), 0.1 (anchors), and 0.1 (titles), again using using the vector space model (`UAmsT07MSum8`) and the language model (`UAmsT07MSm8L`, not pooled).

3.2 Results

The topic set contains 10,000 topics numbered 1 to 10000. Table 1 (top half) shows statistics of the number of judged and relevant documents, based on the "prels" files released on October 1st. In total 1,692 different topics have been assessed. The number of relevant documents per topic varies from 1 to 52, with a mean of 5 and a median 3. For no less than 237 topics, no relevant document has been found. The topic set also includes the adhoc topics of the Terabyte (TB) tracks at TREC 2004-2006. For comparison, we also show their statistics in Table 1 (bottom half). During the three years of the Terabyte track 149 topics have been assessed, with 4 to 617 relevant documents (mean 181 and median 130). There are striking differences between the two sets of judgments: First, the number of topics assessed at the MQ track is roughly ten times larger than the three year of TB track together. Second, the number of judged documents, as well as the number of relevant documents per topic is over ten times larger for the TB topics.

Table 2: Results for the MQ track.

	Million Query		Terabyte 2004-2006		
`UAmsT07`	NEU	UMass	map	bpref	P@10
`...MTeVS`	0.1822	85.13	0.1654	0.2527	0.3047
`...MTeLM`	0.2986	–	0.2921	0.3410	0.5376
`...MTiVS`	0.0902	–	0.0369	0.0939	0.2168
`...MTiLM`	0.0956	47.47	0.0392	0.0977	0.2154
`...MAnVS`	0.0564	–	0.0274	0.0763	0.2081
`...MAnLM`	0.0655	34.66	0.0278	0.0742	0.2034
`...MSum6`	0.1804	93.86	0.1398	0.2348	0.2953
`...MSm6L`	0.2273	–	0.2347	0.3069	0.3738
`...MSum8`	0.2004	98.14	0.1621	0.2482	0.3094
`...MSm8L`	0.2910	–	0.2696	0.3273	0.4711
Topics	1,083	1,692	149	149	149

Table 3: Results for the MQ track using the shallow judgments as qrels.

	Million Query		
`UAmsT07`	map	bpref	P@10
`...MTeVS`	0.1487	0.2584	0.1451
`...MTeLM`	0.2503	0.3472	0.2396
`...MTiVS`	0.0741	0.1586	0.0972
`...MTiLM`	0.0815	0.1537	0.1064
`...MAnVS`	0.0532	0.1228	0.0823
`...MAnLM`	0.0610	0.1205	0.0941
`...MSum6`	0.1545	0.2576	0.1615
`...MSm6L`	0.1906	0.3163	0.1835
`...MSum8`	0.1679	0.2629	0.1671
`...MSm8L`	0.2460	0.3470	0.2330
Topics	1,692	1,692	1,692

Table 2 shows the results for the MQ track. The first two scores are based on the MQ judgments: NEU stands for the estimated MAP (statMAP) as produced by the Northeastern University's method, UMass stands for the expected MAP as produced by the University of Massachusetts Amherst's method.[1] Comparing the scores over the five runs, we see one upset: whereas NEU prefers the full-text vector-space run (TeVS) over the vector-space combination (Sum6), UMass has it the other way around. Both NEU and UMass methods agree on the best of the five runs: the vector-space combination (Sum8). Over all runs, the NEU method gives the highest statMAP score to the full-text language model run (TeLM). The next three scores in Table 2 are based on the TB assessments. The best scoring run on all measures is the full-text language model run (TeLM). The order of the five official submissions is, again, differt: now the full-text vector-space run (TeVS) scores best.

What if we treat the MQ judgments as as normal qrels (so assuming that non-judged documents are non-relevant)? Table 3 shows the results. The best scoring run, again on

[1] We failed to reproduce the "official" scores, and hence only include these for the five official runs.

Table 4: Rank correlations of the resulting system rankings (columns and rows are in the same order).

Million Query		Terabyte			Million Query		
NEU	UMass*	map	bpref	P@10	map	bpref	P@10
–	0.800	0.956	0.911	0.911	0.956	0.911	0.956
–	–	0.600	0.600	0.800	1.000	0.800	1.000
–	–	–	0.956	0.867	0.911	0.867	0.911
–	–	–	–	0.911	0.867	0.911	0.867
–	–	–	–	–	0.867	1.000	0.867
–	–	–	–	–	–	0.867	1.000
–	–	–	–	–	–	–	0.867
–	–	–	–	–	–	–	–

* Comparisons are restricted to 5 runs.

Table 5: Relevant, nonrelevant, and unjudged documents for MQ judged topics (top) and TB judged topics (bottom).

	Rank	Relevant		Nonrelevant		Unjudged	
		#	%	#	%	#	%
Text	1	588	34.75	704	41.61	400	23.64
	10	4,047	23.92	5,760	34.04	7,044	41.63
	100	12,073	7.14	21,678	12.81	134,283	79.36
Anchors*	1	371	21.93	931	55.02	390	23.05
	10	1,605	9.49	4,909	29.01	10,198	60.27
	100	3,052	1.80	8,837	5.22	153,554	90.75
Text*	1	82	55.03	67	44.97	0	0.00
	10	801	53.76	682	45.77	7	0.47
	100	5,726	38.43	8,039	53.95	1,135	7.62
Anchors	1	52	34.90	56	37.58	41	27.52
	10	302	20.27	545	36.58	643	43.15
	100	1,319	8.85	3,821	25.64	9,643	64.72

* Run was in the pool.

all measures, is the full-text language model run (TeLM). The best official submission is the vector-space combination (Sum8), in agreement with both the NEU and UMass methods. In fact, the five official submission get the same order by map and by the expected MAP of the UMass method. More generally, we see that map and precision at 10 are resulting in the same system ranking, and that the precision at 10 scores are much lower than for the TB topics in Table 2. This is a clear indication that we have only unearthed a small sample of the relevant documents.

We have now shown three "qrels" and eight measures, how do these agree? Table 4 shows Kendall's tau of the system rank correlation. Some observations present themselves: First, we see that there is reasonable correlation between all pairs of measures, with correlations ranging from 0.6 to 1.0, with the 0.6 for the agreement between UMass and TB map, and UMass and TB bpref. Second, the agreement between NEU and UMass—both calculating 'true' MAP based on the judged sample—is relatively low with 0.8.

What is the impact of low pooling depth? We look at the number of relevant, nonrelevant, and unjudged documents in runs both inside and outside of the judgment pools. The results are shown in Table 5. Looking at the 1,692 MQ topics, over 20% of the top 1 results have not been judged. At rank 10, the percentage of unjudged documents is 42% (full-text, not pooled) and 60% (anchor-texts, pooled). The relative precision over judged documents is still 70% (full-text) and 33% (anchor-texts) suggesting strongly that the judgments are merely a sample. A clear call for caution to use the MQ judgments as traditional qrels (as is done in Table 3). For the MQ topics we see no significant difference between the coverages of runs in and outside the pools. In a sense this may make the comparison of official and post-submission runs less unfair. Looking at the 149 TB topics, we see clearly the difference in the percentage of judged documents for the pooled run (full-text, very similar runs were in the top 50 pools at TREC 2004-2006), and outside the pool (anchor-texts).

4 Conclusions

During the TREC 2007 Million Query track, we submitted a number of runs using different document representations (such as full-text, title-fields, or incoming anchor-texts), and compared results of the earlier Terabyte tracks to the Million Query track. The initial results show broad agreement in system rankings over various measures on topic sets judged at both Terabyte and Million Query tracks, with runs using the full-text index giving superior results on all measures, but also some noteworthy upsets.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO, grants # 612.066.513, 639.072.601, and 640.001.501), and by the E.U.'s 6th FP for RTD (project MultiMATCH contract IST-033104).

References

[1] ILPS. The ILPS extension of the Lucene search engine, 2007. http://ilps.science.uva.nl/Resources/.

[2] J. Kamps. Effective smoothing for a terabyte of text. In *The Fourteenth Text REtrieval Conference (TREC 2005)*. National Institute of Standards and Technology. NIST Special Publication, 2006.

[3] J. Kamps. Experiments with document and query representations for a terabyte of text. In *The Fifteenth Text REtrieval Conference (TREC 2006)*. National Institute of Standards and Technology. NIST Special Publication, 2007.

[4] J. Kamps, C. Monz, M. de Rijke, and B. Sigurbjörnsson. Approaches to robust and web retrieval. In *The Twelfth Text REtrieval Conference (TREC 2003)*, pages 594–599. National Institute of Standards and Technology. NIST Special Publication 500-255, 2004.

[5] J. Kamps, G. Mishne, and M. de Rijke. Language models for searching in Web corpora. In *The Thirteenth Text REtrieval Conference (TREC 2004)*. National Institute of Standards and Technology. NIST Special Publication 500-261, 2005.

[6] Lucene. The Lucene search engine, 2007. http://lucene.apache.org/.

[7] Snowball. Stemming algorithms for use in information retrieval, 2007. http://www.snowball.tartarus.org/.