The University of Melbourne in the Million Query Track of TREC 2007

William Webber Vo Ngoc Anh Alistair Moffat

Department of Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia

{wew,vo,alistair}@csse.unimelb.edu.au

October 17, 2007

Runs

The University of Melbourne submitted runs from four systems to the TREC 2007 Million Query track, as summarized in Table 1.

The `umelbstd` system used the Zettair information retrieval engine (`http://seg.rmit.edu.au/zettair`), with the default similarity measure, which is based on a language model with Dirichlet priors. Neither stemming nor stopping was employed. The Zettair engine uses a dynamic query pruning technique to reduce query time, as described in Lester et al. [2005]. For the `umelbexp` runs, each query was first posed to a commercial search engine, with the domain for the query restricted to the `.gov` domain. The query was then expanded with the non-stop words contained in the first five snippets, and the expanded query was submitted to the same system as `umelbstd`.

The `umelbimp` system used the impact-based system described by Anh and Moffat [2005]. This approach is based on the vector space model, where each distinct term in the document collection is categorised as a dimension. An object (document or query) is represented as an impact vector in this space, where the projection of the object in each dimension is called the impact of the corresponding term in this object and is defined as an integral number valued between

`umelbstd`	Language model with Dirichlet priors
`umelbexp`	Expand query with snippets from public search engine, then `umelbstd`
`umelbimp`	Impact-based vector space model
`umelbsim`	Merge results from `umelbimp` and `umelbstd`

Table 1: Summary of systems participating in the Million Query track.

System	Score			
	MAP	RBP $p = 0.95$	staMAP	expMAP
`umelbstd`	0.2532	0.6167 +0.0335	0.2663	148.3
`umelbexp`	0.1251	0.3284 +0.4527	0.1474	95.9
`umelbimp`	0.2499	–	0.2621	147.1
`umelbsim`	0.2641	–	0.2967	170.6

Table 2: Summary results. The metrics MAP and RBP (rank biased precision) are scored against the 149 topics drawn from the preceding three years' Terabyte tracks. The staMAP measure estimates MAP via the sampling approach proposed by NEU. The expMAP measure is as reported by UMass's mtc system.

0 and 8 inclusively. The similarity score between a document and a query is the scalar product of corresponding vectors, and is also integral.

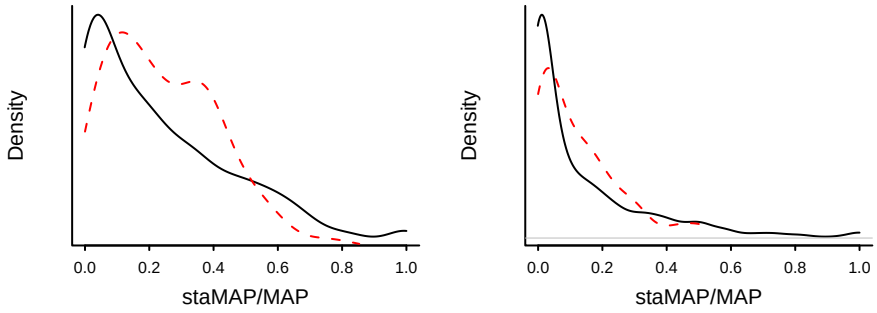
The `umelbsim` system resulted from merging the outputs of `umelbstd` and `umelbimp`. The intention was simple: the two original runs employ different similarity measures, and, in this case, even different parsing policies, hence merging might help to improve the effectiveness. To avoid the complication caused by the different nature of the similarity scores (`umelbstd` gives floating point scores with almost no ties, while `umelbimp` gives integral scores with a great chance of ties), the merging was based on rank rather than on actual score. The new score s_d for a document d is defined as

$$s_d = (\max +2 - r_d^1) \times (\max +2 - r_d^2)$$

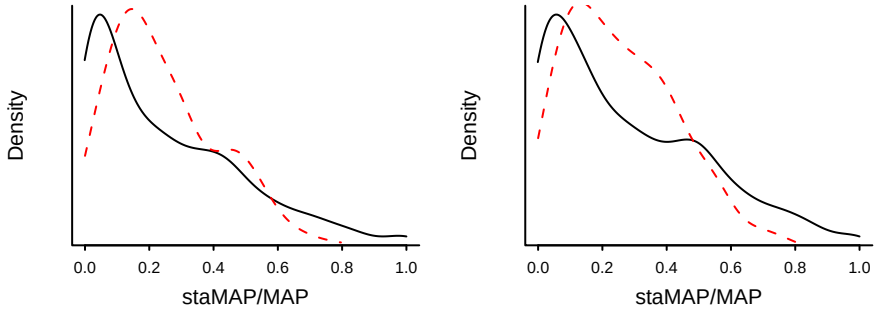
where r_d^1 and r_d^2 are the ranks of d in the first and second runs, respectively, and $\max = 1000$ is the maximal rank. If a document does not appear in the result list of a run, its rank in that run is taken to be $\max +1$.

Results

Table 2 gives the results as provided by NIST. The MAP scores are from topics (and relevance judgments) provided in the previous three years of the Terabyte track. Documents for judging were selected in these years via pooling, although the exact method differed slightly from year to year. The judgments are therefore are "full", but only for the originally pooled systems. The runs submitted to the current track, however, may have many high-ranking documents unjudged. There is no way to tell from the MAP scores how significant these unjudged documents are; they are simply assumed to be irrelevant. In contrast, the rank biased precision (RBP) metric [Moffat et al., 2007] quantifies the degree of uncertainty via an error residual, reported after the plus sign in Table 2. It can immediately be observed, for instance, that the `umelbexp` has returned a large number of unjudged documents, almost half of the documents, as weighted by RBP. In this case, it would not be surprising if most of the new documents are in fact irrelevant, as `umelbexp` was very much an ad-hoc experiment in doing something "different". Nevertheless, the RBP score alerts us to the possibility that the system is being unfairly penalised. A paired, two-tailed t-test on MAP scores finds significant difference between `umelbexp` and every other system at level $p = 1e-16$;



(a) `umelbstd` (b) `umelbexp`



(a) `umelbimp` (b) `umelbsim`

Figure 1: Gaussian kernel density estimates for staMAP (solid black line) and MAP (dashed red line) for each of the submitted systems.

`umelbsim` is significantly different from both `umelbimp` and `umelbstd` at level $p = 0.05$; but the difference between `umelbstd` and `umelbimp` is not significant. It therefore appears that the merging of the `umelbimp` and `umelbstd` into `umelbsim` has produced a composite system better than its components.

The staMAP and expMAP figures reported in Table 2 are estimations of MAP using methods proposed by NEU [Yilmaz and Aslam, 2006] and UMass; the figures as provided by NIST have been reported. Evidently, there is something wrong with the scale of the expMAP figures. Nevertheless, the two estimation techniques agree in ordering and relative magnitude with the "full" MAP scores, and the estimates provided by staMAP are surprisingly close to the full MAP scores – surprising, because they are based on an entirely different set of topics.

Figure 1 explores the distribution of MAP and staMAP scores for each of the submitted runs. Note that, oddly enough, staMAP produced a number of MAP estimates that were greater than 1.0; these have been omitted. Additionally, topics for which there were no relevant documents have also been excised. In each case, the MAP score observed on the historical TREC topic sets are less dispersed than the estimates produced by staMAP for the new topics, which were sampled from a real-world query log. This effect might be due to the conscious effort placed in the past into designing TREC topics that are neither too easy nor too hard, or it could be an artifact of the estimation technique in staMAP itself.

The minimal test collection or *mtc* method of UMass [Carterette, 2007] is based around esti-

System A	System B		
	umelbexp	umelbimp	umelbsim
umelbstd	52.463	1.181	−22.305
	0.837	0.725	0.394
umelbexp	–	−51.282	−74.758
		0.885	0.907
umelbimp	–	–	−23.486
			0.414

Table 3: Estimated MAP deltas (top) and delta variances (bottom) from the mtc method. The estimated delta is for System A compared to System B; positive deltas mean that System A is superior to System B.

mating probabilities of relevance for unjudged documents, and is aimed primarily at discriminating between pairs of systems. It produces not just an estimate of the difference in MAP scores between systems, but also a variance on that estimate and a confidence that the actual difference that would result from full evaluation is non-zero. These figures were provided by NIST, and are reported in Table 3. The mtc method gives a probability of 1.0 that full-evaluation deltas are non-zero for every pairing except for `umelbstd` and `umelbimp`, where the probability is given as 0.92. These results accord with those reported in Table 2, although obviously once again there is a problem with scaling.

Judging

The University was also actively involved in performing relevance judgments, judging a total of 28 topics. Some observations upon the judging experience might be germane.

First, on the question of topic selection. Topics were chosen at random from the full query set, which was derived from a real-world query log. The assessor was then presented with a list of ten randomly selected topics, and asked to choose one to actually assess. However, the assessor's choice amongst these ten topics is far from random. The experience of the Unversity's assessor was that the underlying information need was not readily comprehensible from the query itself in most cases. The assessor is therefore constrained to choose topics that they themselves understand; and even amongst these topics, there is a natural tendency to choose the topic that is clearest. What effect this bias has upon the makeup of topics that made it into the final topic set is a matter for speculation.

Second, there is the issue of the document corpus. The Million Query track used the same `gov2` collection as the Terabyte track of TREC has been using. On each occasion that this document corpus is used, one is reminded once again of the enormous volume of essentially identical pages that occur in it multiple times; Zobel and Bernstein [2006] place the proportion of redundant documents at over 25% of the collection. Serious thought needs to be given to either purging these redundant documents from the collection, or else systematically ensuring that relevance judgments made for one document in an equivalence set are carried through to the rest.

References

Vo Ngoc Anh and Alistair Moffat. Simplified similarity scoring using term ranks. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *Proc. 28th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 226–233, Salvador, Brazil, August 2005.

B. Carterette. Robust test collections for retrieval evaluation. In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Amsterdam, July 2007.

N. Lester, A. Moffat, W. Webber, and J. Zobel. Space-limited ranked query evaluation using adaptive pruning. In Anne H.H. Ngu, Masaru Kitsuregawa, Erich J. Neuhold, Jen-Yao Chung, and Quan Z. Sheng, editors, *Proc. 6th Int. Conf. on Web Informations Systems*, pages 470–477, New York, November 2005. LNCS 3806, Springer.

Alistair Moffat, William Webber, and Justin Zobel. Strategic system comparisons via targeted relevance judgments. In C. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. de Vries, editors, *Proc. 30th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Amsterdam, July 2007.

Emine Yilmaz and Javed A. Aslam. Estimating average precision with incomplete and imperfect judgments. In Philip S. Yu, Vassilis Tsotras, Edward Fox, and Bing Liu, editors, *Proc. 15th ACM Int. Conf. on Information and Knowledge Management*, pages 102–111, Arlington, Virginia, USA, August 2006.

J. Zobel and Y. Bernstein. The case of the duplicate documents: Measurement, search, and science. In X. Zhao, J. Li, H.T. Shen, M. Kitsuregawa, and Y. Zhang, editors, *Proc. APWeb Asia Pacific Web Conference*, pages 26–39, Harbin, China, January 2006. LNCS 3841.