Fondazione Ugo Bordoni at TREC 2004

Giambattista Amati, Claudio Carpineto, and Giovanni Romano

Fondazione Ugo Bordoni
Rome Italy

Abstract

Our participation in TREC 2004 aims to extend and improve the use of the DFR (Divergence From Randomness) models with Query Expansion (QE) for the robust track. We experiment with a new *parameter-free* version of Rocchio's Query Expansion, and use the information theory based function, Info_{DFR} to predict the AP (Average Precision) of queries. We also study how the use of an external collection affects the retrieval-performance.

1 Introduction

FUB participation in the robust track deals with the adaptation of the DFR modular probabilistic framework[2, 4, 1, 3] together with query expansion based on distribution analysis[5, 6, 1].

In the robust track of TREC 12 [10] several approaches used external collections for the term-weighting [7, 11], with the aim of enhancing both MAP (Mean Average Precision) and performance over poor-performing queries (e.g. MAP(X)). Therefore, we want to measure the impact of exploiting external resources in retrieving documents of the target collection.

All runs employ a new parameter-free model of QE that is presented in Section 2.3. We use a unique query-performance predictor, which is displayed in Equation 8. For each baseline run (fub04Te, fub04De and fub04TDNe) we add an additional QE processing in their analogue runs (fub04Tge, fub04T2ge, fub04Dge and fub04TDNge) before applying the final QE. More precisely, we first expand the initial query q (T, D or TDN) with the same methodology used for the baselines (and explained in Section 2.3) using the first **R** retrieved documents of the

external collection, and thus obtain a new expanded query q_1. Then, we submit q_1 as a new query to the target collection. This time, we retrieve **R** documents of the target collection. We expand the query q_1 with the same technique used for the baselines, and obtain the final ranking. For fub04Tge, fub04Dge and fub04TDNge we use a sort of oracle (a WEB search engine) to select the documents of the external collection, while we exploit the .GOV collection for the run fub04T2ge. In the runs fub04Tg, fub04Dg and fub04TDNg we do not apply a second pass QE on the target collection, but we directly retrieve documents of the target collection with the expanded query q_1.

However, we have only submitted the title field for the longest queries (TDN) in the runs using the WEB search engine for the first-pass retrieval for the query-expansion. Similarly, we extract the most significant terms from for the only-description queries using the within-query term-weights of Equation 4 for the runs fub04Dge and fub04Dg. As for the longest queries, the title is submitted as it is for the query-expansion in the remaining runs fub04Tge and fub04Tg.

2 Term-weighting models

We use only one DFR within-document term-weighting formulas and only one model of query expansion. The term-weighting model is I(n)OL2:

$$\mathbf{weight}(\mathbf{t}|\mathbf{d}) = \frac{\mathbf{tfn}}{\mathbf{tfn}+1}\log_2\left(\frac{|\mathbf{D}|-\mathbf{df}+1}{\mathbf{df}+0.5}\right) \tag{1}$$

$$\text{where} \quad \begin{cases} \mathbf{tfn} = & \text{tf}(\mathbf{t}|\mathbf{d})\cdot\log_2\left(1+c\cdot\dfrac{\mathbf{adl}}{\mathbf{dl}}\right) \\ |\mathbf{D}| & \text{is the size of the collection } \mathbf{D} \\ \mathbf{df} & \text{the document-frequency} \\ \mathbf{dl} & \text{the document-length} \\ \mathbf{adl} & \text{the average document-length} \end{cases} \tag{2}$$

The value of the parameter c of the within-document term-weighting DFR models is set to 2 [4, 1, 2, 3]. I(n)OL2 can be seen as a generalization of the $BM25$ formula[9]. To see this, we introduce a new parameter k_1 (however, we have used $k_1 = 1$ in all our experiments):

$$\frac{\mathbf{tfn}}{\mathbf{tfn}+k_1}\log_2\left(\frac{|\mathbf{D}|-\mathbf{df}+1}{\mathbf{df}+0.5}\right)$$

and denote $\dfrac{\mathbf{dl}}{c\cdot\mathbf{adl}}$ by the variable x. Then:

$$\frac{\mathbf{tfn}}{\mathbf{tfn}+k_1} = \frac{\text{tf}(\mathbf{t}|\mathbf{d})}{\text{tf}(\mathbf{t}|\mathbf{d}) + \dfrac{k_1}{\log_2(x+1) - \log_2 x}}$$

The Taylor series expansion at the point $x = 1$ of the function

$$\frac{k_1}{\log_2(x+1) - \log_2 x}$$

with error $O((\dfrac{\mathbf{dl}}{c \cdot \mathbf{adl}} - 1)^3)$ is:

$$k_1 \cdot \left(1 + \log_2 e \cdot 0.5 \cdot (\frac{\mathbf{dl}}{c \cdot \mathbf{adl}} - 1) - \frac{1}{8}\log_2 e \cdot (3 - 2\log_2 e)(\frac{\mathbf{dl}}{c \cdot \mathbf{adl}} - 1)^2\right) =$$
$$= k_1 \cdot \left(0.2580 + 0.7627 \cdot \frac{\mathbf{dl}}{c \cdot \mathbf{adl}} - 0.0207 \cdot \frac{\mathbf{dl}}{(c \cdot \mathbf{adl})}^2\right)$$

If k_1 is set to 1.2 as the default value of k_1 in the $BM25$ formula, then:

$$\frac{\mathbf{tfn}}{\mathbf{tfn}+k_1} \sim \frac{\text{tf}(\mathbf{t}|\mathbf{d})}{\text{tf}(\mathbf{t}|\mathbf{d}) + 0.3096 + 0.9152 \cdot \frac{\mathbf{dl}}{c \cdot \mathbf{adl}}}$$

that is when $c = 1$, I(n)OL2 formula is approximated by the $BM25$ formula. In its more general form, I(n)OL2 is a generalization of $BM25$ (In Terrier [8] I(n)OL2 is called DFR_BM25).

2.1 Query expansion

The QE method is a parameter-free extension of that used in TREC 2003 [3]. We have evaluated this methodology with all other TREC collections achieving often better results than with the optimal Rocchio parameter β.

2.2 Old model of QE

The last TREC within-query term-weights of the expanded query q^* of the original query q was obtained as follows:

$$\text{weight}(\mathbf{t} \in q^*) = \quad \mathbf{ntfq} + \beta \cdot \frac{\text{Info}_{\text{DFR}}(\mathbf{t})}{\text{MaxInfo}_{\text{DFR}}} \tag{3}$$

where:

$$\begin{cases} \mathbf{ntfq} = & \dfrac{\mathbf{tfq}}{\arg\max_{t \in q} \mathbf{tfq}} \\ \mathbf{tfq} & \text{is the within-query term-frequency } \mathbf{tfq} \text{ of the term} \\ \text{Info}_{\text{DFR}}(\mathbf{t}) = & -\log_2 \text{ Prob}(\text{ tf}(\mathbf{t}|\mathbf{R})|\text{ tf}(\mathbf{t}|\mathbf{D})) \\ \text{MaxInfo}_{\text{DFR}} = & \arg\max_{\mathbf{t} \in q^*} \text{Info}_{\text{DFR}}(\mathbf{t}) \\ \text{tf}(\mathbf{t}|\mathbf{R}) & \text{is the term-frequency in the pseudo-relevant set } \mathbf{R} \\ \text{tf}(\mathbf{t}|\mathbf{D}) & \text{is the term-frequency in the collection} \\ \text{Prob} & \text{is the probability of term-frequency computed by any DFR model} \\ \beta & \text{is a parameter in the interval } [0,1] \text{ (last year it was } 0.4). \end{cases} \quad (4)$$

In particular, we use the following Bose-Einstein statistics (Bo2) as DFR model:

$$\begin{aligned} \text{Info}_{\text{Bo2}}(\mathbf{t}) = & \quad -\log_2\left(\tfrac{1}{1+\lambda}\right) - \text{ tf}(\mathbf{t}|\mathbf{R}) \cdot \log_2\left(\tfrac{\lambda}{1+\lambda}\right) \quad \text{[Bo2]} \\ \lambda = & \quad \sum_{\text{t'}} \text{tf}(\text{t'}|\mathbf{R}) \cdot \frac{\text{tf}(\mathbf{t}|\mathbf{D})}{\sum_{\text{t'}} \text{tf}(\text{t'}|\mathbf{D})} \end{aligned} \quad (5)$$

where **R** denotes the pseudo-relevant set.[1]

2.3 New model of QE: a parameter-free expanded query

An upper bound of $\text{Info}_{\text{DFR}}(\mathbf{t})$ is when the divergence of $\text{tf}(\mathbf{t}|\mathbf{D})$ and $\text{tf}(\mathbf{t}|\mathbf{R})$ is maximum, that is when $\text{tf}(\mathbf{t}|\mathbf{R}) = \text{tf}(\mathbf{t}|\mathbf{D})$. Let **M** be the maximum value of all values

$$\begin{aligned} \mathbf{M} &= \arg\max_{\mathbf{t} \in q^*} \mathbf{M}(\text{t}) \\ \mathbf{M}(\text{t}) &= \lim_{\text{tf}(\mathbf{t}|\mathbf{D}) \to \text{tf}(\mathbf{t}|\mathbf{R})} \text{Info}_{\text{DFR}}(\mathbf{t}) \end{aligned} \quad (6)$$

In different words $\mathbf{M}(\text{t})$ is obtained by substituting in Info_{DFR} each occurrence of $\text{tf}(\mathbf{t}|\mathbf{D})$ with its theoretical lower bound $\text{tf}(\mathbf{t}|\mathbf{R})$.

The new within-query term-weight in the expanded query q^* of the original query q is obtained as follows:

$$\text{weight}(\mathbf{t} \in q^*) = \quad \mathbf{ntfq} + \frac{\text{Info}_{\text{DFR}}(\mathbf{t})}{\mathbf{M}} \quad (7)$$

[1] A new query-term must also appear at least in 2 retrieved documents. This condition is to avoid the noise of the highly informative terms which appear only once in the set of the topmost retrieved documents.

In the implementation we select the term $\mathbf{t}$ with the highest $\text{Info}_{\text{DFR}}(\mathbf{t})$, which in general has also the highest term frequency $\text{tf}(\mathbf{t}|\mathbf{R})$ in the topmost documents. Then, we compute the $\mathbf{M}(\text{t})$ value as $\mathbf{M}$ associated to this selected terms. This technique corresponds to anautomatic way to select a query-biased value for the parameter β.

2.4 Parameters

The only parameters we use in QE are:

- $|\mathbf{R}| = 8$. [2]
- the number of terms $\mathbf{N}_\text{t}$ of the expanded query is 40.[3]

3 Predicting topic difficulty with Info $_{\text{DFR}}$

Our notion of query-difficulty is based on the same within-query term-weighting Info $_{\text{DFR}}$ gained after a first-pass ranking. If there is a significant divergence in the query-term frequencies before and after the retrieval, then we make the hypothesis that this divergence is caused by a query which is easy-defined.

$$\text{difficulty score of } q \ =_{\text{def}} \ \sum_{\mathbf{t} \in q} \text{Info}_{\text{DFR}}(\mathbf{t}) \text{ of Formula 4} \tag{8}$$

where DFR is a basic model (based on the Binomial, the Bose-Einstein statistics or the Kullback-Leibler divergence measure). We use the probability of Bose-Einstein as defined in Formula (5) in all our runs and in all QE processes.

To compute the difficulty-score of the query we first produced a first-pass ranking as it is done in QE. We took the set $\mathbf{R}$ of the first 8 retrieved documents and we compute the score $\text{Info}_{\text{DFR}}(\mathbf{t})$ for each term occurring in the query. We consider the query-terms which appear at least twice in these pseudo-relevant documents. This score reflects the amount of information carried by the query-term within these pseudo-relevant documents.

[2] This parameter can be also eliminated query-by-query by choosing the number of topmost retrieved documents which maximise a normalised version of our (or any other) query-performance predictor.

[3] This parameter is the least important among the QE parameters. Indeed, even using many more expanded terms the performance does not change significantly.

Table 1: Mean Average Precision and number of topics with no relevant document in the top 10 retrieved documents of all ten runs.

	old queries		new queries		hard queries		all queries	
runs	MAP	Nr top.	MAP	Nr top.	MAP	Nr top.	MAP	Nr top.
fub04T2ge	0.2870	17.0%	0.3295	16.3%	0.1331	26.0%	0.2954	16.9%
fub04Te	0.2881	17.0%	0.3322	12.2%	0.1367	26.0%	0.2968	16.1%
fub04Tg	0.2869	10.0%	0.3464	10.2%	0.1388	18.0%	0.2986	10.0%
fub04Tge	0.2985	13.0%	0.3514	12.2%	0.1450	22.0%	0.3089	12.9%
fub04De	0.2890	15.0%	0.3760	10.2%	0.1374	24.0%	0.3062	14.1%
fub04Dg	0.2954	7.5%	0.3636	8.2%	0.1470	14.0%	0.3088	7.6%
fub04Dge	0.3093	9.5%	0.3823	8.2%	0.1475	18.0%	0.3237	9.2%
fub04TDNe	0.3285	8.0%	0.3824	6.1%	0.1780	14.0%	0.3391	7.6%
fub04TDNg	0.3171	6.5%	0.3635	8.2%	0.1668	10.0%	0.3262	6.8%
fub04TDNge	0.3323	9.0%	0.3741	8.2%	0.1775	18.0%	0.3405	8.8%

4 Results

We submitted 4 title-only (the runs with the descriptor T), 3 description-only (the runs with the descriptor D), and 3 full-query runs (the runs with the descriptor TDN). Results are shown in Tables 1 and 2. We remind that the runs whose names contain the label "g" or "ge" are those that use external collections in a first-pass retrieval. It is a matter of fact that the support of a very large additional collection, such as the whole collection of the WEB documents, improves the retrieval quality. This is achieved by further expanding the queries, already expanded with the external collection resource, with a second pseudo-relevance feedback over the target collection. This technique works independently from how long the query is obtained with the first loop of the QE process. On the other hand, the additional WEB collection resource seems to be less important or even superfluous with very long queries (TDN). In this case, the pseudo-relevance feedback over the target collection can be sufficient to obtain the maximal performance achievable with the system and the target collection.

The evaluation of the parameter-free model of QE achieves often better results than the model based on the optimal Rocchio parameter β. This can happen because we compute a sort of optimal value for β query-by-query, whilst the standard approach looks for the best-match value of the parameter β for all queries.

Table 2: Kendall measure between AP ranks and query-difficulty ranks. Actual MAP, MAP without the predicted X worst queries and MAP without the actual X worst queries. fub04T2ge employs a random query-performance predictor.

runs	Kendall	MAP	MAP -X =50 pred	MAP -X=50 optimal
fub04T2ge	0.0070	0.2954	0.3064	0.3634
fub04Te	0.2770	0.2968	0.3273	0.3648
fub04Tg	0.2860	0.2986	0.3297	0.3648
fub04Tge	0.2840	0.2985	0.3404	0.3768
fub04De	0.3300	0.3089	0.3514	0.3781
fub04Dg	0.3090	0.3062	0.3456	0.3739
fub04Dge	0.2980	0.3088	0.3625	0.3934
fub04TDNe	0.2640	0.3391	0.3680	0.4114
fub04TDNg	0.3270	0.3262	0.3636	0.3945
fub04TDNge	0.2950	0.3405	0.3774	0.4116

As for the results of Table 2, we also observe, but the results are not reported here, that Kendall's correlation factor is significantly lower than Spearman's correlation coefficient. Moreover, both Kendall's correlation factor and Spearman's correlation coefficient say that the Equation 8 provides a better query-performance prediction when QE is not applied. The reason is that, if there are terms which have a high divergence between the term-frequencies within the set of pseudo-relevant documents and within the collection but are not included in the original query, then the difficult-score is lower than that if they were included. Obviously, these terms are included in the expanded query in the second-pass retrieval, and therefore the quality of the ranking is improved with respect to that of the the initial prediction.

The advantage of our predictor in Equation 8 is that has no additional computational cost if QE is adopted. Table 2 shows that if we eliminate 20% of all queries Equation 8 is able to raise the MAP of an average of 11.15% per run, and with an average of 29.67% achieved by Kendall's correlation factor.

Acknowledgments

The experiments were conducted using Terrier's Information Retrieval platform[4]. Terrier version 1.0.0 implements the parameter-free query expansion method described in Section 2.3 by default. Under the same setting of the run fub04Te (with c=2, $\mathbf{N}_t = 40$, $|\mathbf{R}| = 8$), the last release of Terrier (version 1.0.0) gives better results than the official TREC results. The model used in this set of experiments is called DFR_BM25 in Terrier version 1.0.0. MAP of DFR_BM25 is 0.3015 with Terrier 1.0.0, if the model of query expansion is Bo2, while MAP is 0.3035, if the model of query expansion is Bo1.

References

[1] Giambattista Amati. *Probability Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, University of Glasgow, June 2003.

[2] Gianni Amati, Claudio Carpineto, and Giovanni Romano. FUB at TREC 10 web track: a probabilistic framework for topic relevance term weighting. In E.M. Voorhees and D.K. Harman, editors, *In Proceedings of the 10th Text Retrieval Conference TREC 2001*, pages 182–191, Gaithersburg, MD, 2002. NIST Special Pubblication 500-250.

[3] Gianni Amati, Claudio Carpineto, and Giovanni Romano. FUB at TREC 2003: Robust and web track. In E.M. Voorhees and D.K. Harman, editors, *In Proceedings of the 12th Text Retrieval Conference TREC 2003*, pages 234–245, Gaithersburg, MD, 2004. NIST Special Pubblication 500-255.

[4] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[5] C. Carpineto, R. De Mori, G. Romano, and B. Bigi. An information theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27, 2001.

[6] C. Carpineto, G. Romano, and V. Giannini. Improving retrieval feedback with multiple term-ranking function combination. *ACM Transactions on Information Systems*, 20(3):259–290, 2002.

[4] http://ir.dcs.gla.ac.uk/terrier/index.html

[7] L. Grunfeld, K.L. Kwok, N. Dinstl, and P. Deng. TREC2003 Robust, HARD and QA track experimants using PIRCS. In E.M. Voorhees and D.K. Harman, editors, *In Proceedings of the 12th Text Retrieval Conference TREC 2003*, pages 510–521, Gaithersburg, MD, 2004. NIST Special Pubblication 500-255.

[8] I. Ounis, G. Amati, Plachouras V., B. He, C. Macdonald, and D. Johnson. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on IR Research (ECIR 2005)*, Lecture Notes in Computer Science. Springer, 2005.

[9] S. Robertson, S. Walker, M.M. Beaulieu, M. Gatford, and A. Payne. Okapi at trec-4. In D.K. Harman, editor, *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)*. Department of Commerce, National Institute of Standards and Technology, 1996.

[10] Ellen M. Voorhees. Overview of the trec 2003 robust retrieval track. In E.M. Voorhees and D.K. Harman, editors, *In Proceedings of the 12th Text Retrieval Conference TREC 2003*, pages 69–77, Gaithersburg, MD, 2004. NIST Special Pubblication 500-255.

[11] David L. Yeung, Charles L.A. Clarke, Gordon V. Cormack, Thomas R. Lynam, and Egidio L. Terra. Task-Specific Query Expansion (MultiTest Experiments for TREC 2003. In E.M. Voorhees and D.K. Harman, editors, *In Proceedings of the 12th Text Retrieval Conference TREC 2003*, pages 810–819, Gaithersburg, MD, 2004. NIST Special Pubblication 500-255.