

Fondazione Ugo Bordoni at TREC 2003: robust and web track

Notebook draft

Giambattista Amati, Claudio Carpineto, and Giovanni Romano

Fondazione Ugo Bordoni
Rome Italy

Abstract

Our participation in TREC 2003 aims to adapt the use of the DFR (Divergence From Randomness) models with Query Expansion (QE) to the robust track and the topic distillation task of the Web track. We focus on the robust track, where the utilization of QE improves the global performance but hurts the performance on the worst topics. In particular, we study the problem of the selective application of the query expansion. We define two information theoretic functions which are shown to be correlated respectively with the average precision and with the increase of average precision under the application of QE. The use of the second measure to selectively apply QE allows to achieve a performance comparable to that of the unexpanded method on the worst topics and better performance than full QE on the whole set of topics.

1 Robust Track

FUB participation in the robust track deals with the adaptation of the DFR modular probabilistic framework[2, 3, 1] together with query expansion based on distribution analysis[4, 5, 1] to this new task. In the robust track there are two new evaluation measures, the number of topics with no relevant documents in the top retrieved 10 (denoted by NrNoRel) and MAP(X), a measure related to the average precision of the worst X topics.

Preliminary experiments on the collection against the queries of TREC 6, TREC 7 and TREC 8 showed that:

1. The number of topics with no relevant retrieved in the top 10 ranks, NrNoRel, increased when the QE was activated.
2. Similarly, the MAP of the worst X topics always diminished when QE was adopted.

1.1 Submitted runs: QE was adopted in all queries

At the submission time we did not have a stable and robust QE activation method to improve the performance on both the old and the new evaluation measures. Although this year the description-only queries are quite long, the automatic application of QE to all queries seemed to be the safest way to achieve a higher value of MAP. However, QE is detrimental to both MAP(X) and to NrNoRel measures. We thus submitted 4 description-only runs with full QE to maximize global performance and one description-only run with all unexpanded queries, to partially account for the worst topics.

Our framework was tuned to the new experimental conditions as follows:

1.2 Term-weighting models

We used 4 different DFR within-document term-weighting formulas: I(n)B2, I(n)OL2, I(n_e)OL2, I(n_e)OB2.

The models I(n)OL2, I(n_e)OL2 are variants of the model I(n)L2, while I(n_e)OB2 is a variant of I(n_e)B2.

For sake of space we just report the model I(n)OL2:

$$\begin{aligned} \text{I(n)OL2 : } & \frac{tfn}{tfn + 1} \log_2 \left(\frac{|\text{Collection}| - \text{doc_freq} + 1}{\text{doc_freq} + 0.5} \right) \\ \text{where } tfn = & tf \cdot \log_2 \left(1 + c \cdot \frac{\text{average_document_length}}{\text{document_length}} \right) \end{aligned} \quad (1)$$

The value of the parameter c of the within-document term-weighting DFR models I(n)B2, I(n)OB2, I(n)OL2, I(n_e)OL2 was set to 3 [3, 1, 2].

1.3 Query expansion

The QE method was the same as used in TREC-10 with very good results[2] except for the parameter tuning and some additional expansion weight models.

The weight of a term of the expanded query q^* of the original query q is obtained as follows:

$$\text{weight}(\text{term} \in q^*) = tfq_n + \beta \cdot \frac{\text{Info}_{\text{DFR}}}{\text{MaxInfo}}$$

where

- tfq_n is the normalized term frequency within the original query q (i.e. $\frac{tfq}{\max_{t \in q} tfq}$)
- Info_{DFR} is a term frequency in the expanded query induced by using any DFR model:

$$\begin{aligned} \text{Info}_{\text{DFR}} &= -\log_2 \text{Prob}(\text{term}, \text{TopDocuments} | \text{Collection}) \\ \text{MaxInfo} &= \arg_{t \in q^*} \max \text{Info}_{\text{DFR}} \end{aligned}$$

In particular, for Info_{DFR} we used the normalized Kullback-Leibler measure (KL) [4, 2] and the following Bose-Einstein statistics (Bo2)¹:

$$\begin{aligned} \text{Info}_{\text{Bo2}}(\text{term}) &= -\log_2 \left(\frac{1}{1+\lambda} \right) - \text{Freq}(\text{term} | \text{TopDocuments}) \cdot \log_2 \left(\frac{\lambda}{1+\lambda} \right) \quad [\text{Bo2}] \\ \lambda &= \frac{\text{TotFreq}(\text{TopDocuments}) \cdot \text{Freq}(\text{term} | \text{Collection})}{\text{TotFreq}(\text{Collection})} \end{aligned} \quad (2)$$

where TopDocuments denotes the pseudo-relevant set. The other parameters were set as follows:

- $\beta = 0.4$
- $|\text{TopDocuments}| = 10$ and the number of terms of the expanded query is equal to 40.

1.4 Selective application of QE: new experiments

The official results confirmed the outcomes of our preliminary investigation as shown in Table 1. The unexpanded run achieved the best MAP(X) and the lowest NrNoRel, and the runs with expanded queries achieved the highest values of MAP and precision at 10.

In the following we study the problem of selectively applying QE to the set of topics.

¹The 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.

Table 1: The number of selected documents on the first-pass retrieval is 10, the number of the extracted terms for the query expansion is 40.

Parameters	Runs with QE				Run without QE
	fub03InB2e3	fub03IeOLKe3	fub03InOLe3	fub03IneOLe3	fub03IneOBu3
$c = 3$	DFR Models				
	I(n)B2	I(n.e)OL2	I(n)OL2	I(n.e)OL2	I(n.e)OB2
$\beta = 0.4$	DFR Expansion models				
	Bo2	KL	Bo2	Bo2	-
	old topics				
@10:	0.3360	0.3360	0.3380	0.3300	0.3080
MAP :	0.1317	0.1315	0.1340	0.1343	0.1134
top 10 with No Rel.:	13	12	12	12	7
MAP(X)	0.0047	0.0061	0.0070	0.0057	0.0052
	new topics				
@10 :	0.5000	0.4780	0.4880	0.4660	0.4800
MAP:	0.3552	0.3692	0.3697	0.3614	0.3524
top 10 with No Rel.:	5	6	5	8	4
MAP(X)	0.0192	0.0117	0.0152	0.0098	0.0232
	all topics				
@10:	0.4180	0.4070	0.4130	0.398	0.3940
MAP:	0.2434	0.2503	0.2519	0.2479	0.2329
top 10 with No Rel.	18	18	17	20	11
MAP(X)	0.0084	0.0065	0.0077	0.0058	0.0096

We give two measures Info_{DFR} and InfoQ which are related to the Average Precision (AP) and to the successful application of QE. The Info_{DFR} measure tells us when a topic is *difficult*, in the sense that Info_{DFR} is an indicator of a possible low outcome of AP. InfoQ is instead an indicator for the QE activation.

These findings can be a first step towards the definition of a more stable decision strategy for the selective use of the QE.

1.5 How QE affects the Robust track

Consider as an example the performance of the run fub03InOLe3 in Table 2; fub03InOLe3 uses the model I(n)OL2 (see Formula 1). With full QE, we achieved an increase of MAP equal to +7.5% with respect to the baseline run. If we had an oracle telling us when to apply QE query-by-query, the performance increase would nearly double passing from +7.5% to +13.3%.

However, without the oracle a wrong decision of omitting the QE mechanism would seriously hurt the final MAP of the run. The average gain per query is ~ 0.063 and the gain is much greater than the average loss (~ 0.039). Moreover, the number of cases with a successful application of QE (57 out 100) is larger than the number of the failure cases. Both odds are thus in favour of the application of QE.

In the robust track, the success rate of the QE application was below our expectation. Comparing the figures of Table 2 with those relative to all the 150 queries of the past TREC data, we have observed a detriment of the success rate. The success rate was around 65% with all the 150 old queries of past TREC data. A detriment in precision at 10 was observed for only 15% of all the 150 old queries (against 19% of the TREC 2003 queries).

In addition, the increase of MAP with QE using all the old 150 queries was larger ($\sim +10\%$) than that obtained with the new queries ($\sim +5\%$).

Table 2: Run fub03InOLe3 with description-only topics. The columns with “No QE” contain the number of queries to which the QE was not applied.

Old Topics											
Baseline		run fub03InOLe3 with QE				Runs with the oracle					
MAP	P@10	MAP	%	P@10	%	MAP	%	No QE	P@10	%	No QE
0.1147	0.3100	0.1340	+14.4%	0.3380	+ 8.3%	0.1432	+19.9%	21/50	0.3640	+14.8%	10/50
New Topics											
Baseline		run fub03InOLe3 with QE				Runs with the oracle					
MAP	P@10	MAP	%	P@10	%	MAP	%	No QE	P@10	%	No QE
0.3512	0.4780	0.3697	+5.0%	0.4880	+2.1%	0.3942	+10.9%	22/50	0.5160	+7.4%	9/50
All Topics											
Baseline		run fub03InOLe3 with QE				Runs with the oracle					
MAP	P@10	MAP	%	P@10	%	MAP	%	No QE	P@10	%	No QE
0.2330	0.3940	0.2519	+7.5%	0.4130	+4.6%	0.2687	+ 13.3%	43/100	0.4400	+ 10.5%	19/100

1.6 Selective application of QE: topic difficulty

It is a well known evidence that the QE effectiveness is strictly related to the number of documents which are relevant for a given query in the set of the topmost documents in the ranking. If the early precision of the first-pass retrieval is high, then we have a good chance to extract good additional topic terms together with their relative query-weights. To start our investigation we have first computed the correlation factor between

- the number Rel of relevant documents in the whole collection and the AP value over the 100 queries, and
- between Rel and the precision at 10 (P@10).

The correlation value $-1 \leq \rho \leq 1$ indicates the degree of the linear dependence between the two pair of measurements. When $\rho = 0$ the correlation coefficient indicates that the two variables are independent. When instead there is a linear correlation, the correlation coefficient is either -1 or 1 [7]. A negative correlation factor indicates that the two variables are inversely related.

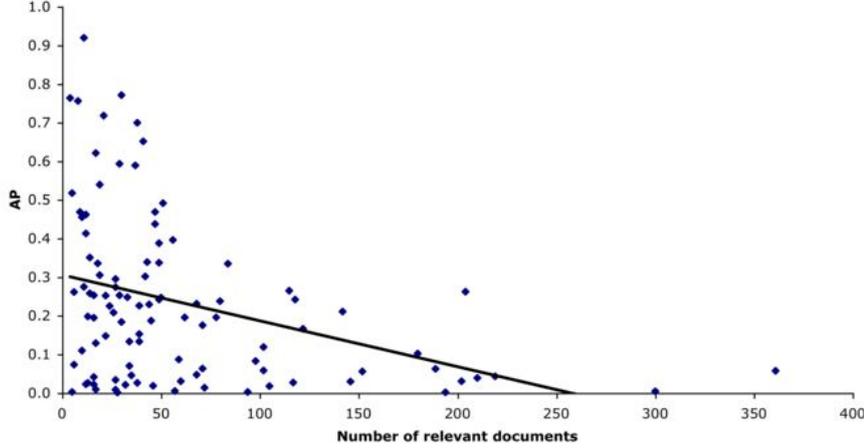
Surprisingly, both these correlation factors come out to be negative ($\rho(Rel, AP) = -0.36$ and $\rho(Rel, P@10) = -0.14$).

Although in these two cases the absolute values of the correlation coefficient are not close to -1 , even small values of the correlation factor are regarded very meaningful especially in large samples [9].

Therefore, these values of the correlation factors seem to demonstrate that the greater the number Rel of relevant documents, the less the precision (MAP and P@10). An approximation line of the scatter line of the AP values for different numbers of relevant documents is produced in Figure 1. The fact that the correlation factor with AP is larger than that with P@10 is due to the definition of AP. The AP measure combines recall and precision by using the number Rel of relevant documents.

This negative correlation might appear to be counter-intuitive, since among the easiest topics there are many which possess a small number of relevant documents, and, as opposite, many difficult topics have many relevant documents. On the other hand, a possible explanation of these negative correlation factors is that a small number of relevant documents for a topic witnesses the fact that the topic is “specific” or “non-general” with respect to the content of the collection. In such a situation, common-sense says that specific queries have few relevant documents, their query-terms have few occurrences in the collection, and they thus are the easiest ones.

Figure 1: The number of relevant documents is inversely related to AP of the unexpanded query ($\rho(Rel, AP) = -0.36$). Queries with many relevant documents contribute little to MAP.



However, a definition of the specificity based on the number of relevant documents for the query would depend on the evaluation; we rather prefer to have a different but operational definition of the query-specificity or query-difficulty.

The notion of query-difficulty is given by the notion of the amount of information Info_{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.

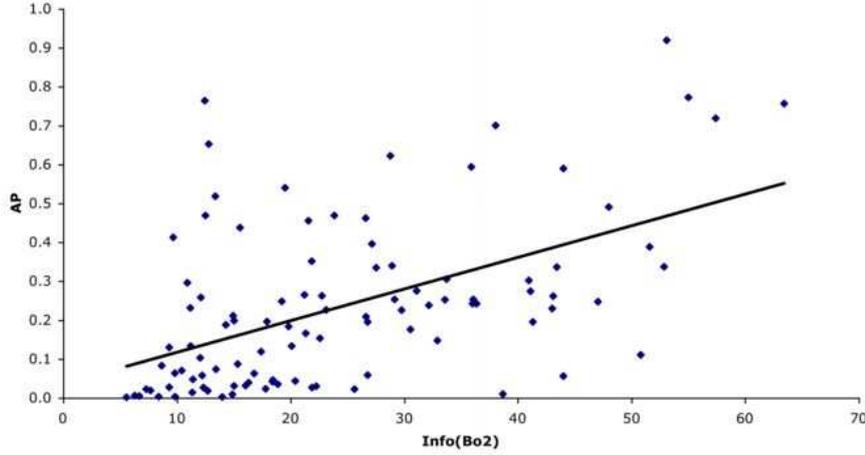
$$\text{Info}_{\text{DFR}} = \sum_{\text{term} \in Q} -\log_2 \text{Prob}(\text{term}, \text{TopDocuments} | \text{Collection}) \quad (3)$$

where $\text{Prob}(\text{term} | \text{Collection}, \text{TopDocuments})$ is a DFR basic model (based on the Binomial, the Bose-Einstein statistics or the Kullback-Leibler divergence measure). We here use the probability of Bose-Einstein as defined in Formula (2). Note that the same weight was used by our expansion method in 3 runs out of the 4 expanded ones (fub03InB2e3, fub03InOLe3, fub03IneOLe3). The Kullback-Leibler divergence was adopted in the run fub03IeOLKe3 (see Table 1).

There are other information theoretic measures capturing the notion of term-specificity of the query. One possible choice, based on the language model, is the clarity score[6], but it is more difficult to implement.

The goodness of Info_{DFR} is tested with the linear correlation factor with AP of the unexpanded queries. The motivation is that easy queries usually yield high AP values. 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 TopDocuments of the first 10 retrieved documents and we computed a score for each term occurring in the query. We considered 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. As shown in Figure 2, Info_{DFR} has a significant correlation with the AP of the unexpanded queries $\rho(\text{Info}_{\text{Bo2}}, AP) = 0.52$. Similarly to the negative correlation between the number of relevant documents and the AP of the unexpanded queries, which is $\rho(Rel, AP) = -0.36$, the correlation factor between the score InfoQ and Rel was negative ($\rho(Rel, \text{Info}_{\text{Bo2}}) = -0.23$). Again, this may be explained by the fact that specific queries possess fewer relevant documents.

Figure 2: The information content Info_{Bo2} of the query within the topmost retrieved documents is linearly correlated to the AP of the unexpanded queries ($\rho(\text{Info}_{\text{Bo2}}, \text{AP}) = 0.52$). Specific queries have a large value of Info_{DFR} .



Unfortunately, we did not find a significant correlation between Info_{DFR} and QE; that is, Info_{DFR} is not able to predict a successful application of QE in a second-pass ranking. These results show that the performance of query expansion is not directly related to query difficult, consistent with the observation [5] that although the retrieval effectiveness of QE in general increases as the query difficult decreases, very easy queries hurt performance.

1.7 Predicting the successful application of QE

Since Info_{DFR} cannot be used as a good indicator for the performance of the QE, we explored alternative information-theoretic functions. The function

$$\text{InfoPriorQ} = \sum_{\text{term} \in Q} -\log_2 \frac{\text{Freq}(\text{term}|\text{Collection})}{\text{TotFreq}(\text{Collection})}$$

was shown to have a moderately weak negative correlation with QE ($\rho(\text{QE}, \text{InfoPriorQ}) = -0.27$). InfoPriorQ is linearly related to the length of the query ($\rho(\text{QueryLength}, \text{InfoPriorQ}) = 0.90$), so that InfoPriorQ does not differ to much from the query length. In other words, the query length is an alternative good indicator for the successful application of the QE. Short queries need in general QE whilst very long queries do not need QE, but this simple fact does not solve the problem of moderately long queries for which QE may or may not succeed.

Let

$$M_Q = \max \left(\frac{\text{InfoPriorQ} - \mu_{\text{InfoPriorQ}}}{\sigma_{\text{InfoPriorQ}}}, \max_{M \in \text{DFR}} \arg \frac{\text{Info}_{\text{DFR}} - \mu_{\text{Info}_{\text{DFR}}}}{\sigma_{\text{Info}_{\text{DFR}}}} \right)$$

The function:

$$\text{InfoQ} = \frac{1}{\text{QueryLength}} \left(\frac{\text{InfoPriorQ} - \mu_{\text{InfoPriorQ}}}{\sigma_{\text{InfoPriorQ}}} + M_Q \right) \quad (4)$$

where the μ_X s and the σ_X s stand for the mean and the standard deviation of the X values, combines InfoPriorQ and Info_{DFR} . Info_{DFR} query rankings may not agree using different DFR models. Because the correlation factor is negative, and since we trigger the QE when InfoQ is

Figure 3: The information content InfoQ of the query based on the combination of the priors and Info_{DFR} within the topmost retrieved documents is negatively correlated to the AP increase with the QE ($\rho(\text{QE increase rate}, \text{InfoQ}) = -0.33$). The first and the third quadrants contain the errors when the threshold is set to 0.

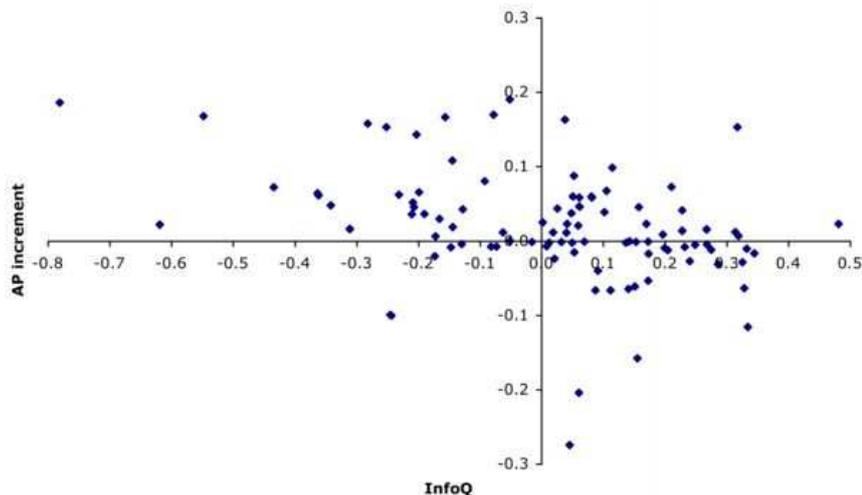


Table 3: The set of queries with the highest InfoQ . The QE is not applied to such queries.

QE success	InfoQ	Query Length	Topic
y	0.482	7	604
n	0.345	8	631
n	0.335	17	320
n	0.333	13	638
n	0.329	9	621
n	0.327	14	619

below a given threshold, a cautious way to smooth different Info_{DFR} values is to compare the threshold to the maximum value of all these DFR models, InfoPriorQ included .

InfoQ has a higher correlation with QE ($\rho(\text{QE}, \text{InfoQ}) = -0.33$) than InfoPriorQ (see Figure 3), and a smaller correlation factor with the query length² ($\rho(\text{QE}, \text{InfoQ}) = 0.62$).

1.8 Conclusions on the selective application of QE

In Table 4 we summarize the results on the selective application of QE. The $\text{MAP}(X)$ values are not reported since the new values are similar to those in the official runs; thus we focus on the other measures. We compare the performance of the 4 submitted runs with QE with the performance of the new runs under the same setting except for the selective application of QE.

The first remark is that the decision rule for QE activation is quite robust. The MAP of the new runs is greater than the MAP of the official runs for a large range of values of the threshold

²Using $\log_2(\text{QueryLength})$ instead of QueryLength the score of Formula 4 is more correlated to the query length with $\rho(\text{QueryLength}, \text{InfoQ}) = 0.74$ and $\rho(\text{QE}, \text{InfoQ}) = -0.34$.

parameter (≥ 0). In fact, InfoQ provides with a high degree of confidence the cases in which QE should be absolutely activated, which are the cases when InfoQ assumes very small negative values, as it can be seen in Figure 3. This explains why the new value of MAP keeps constantly larger than the MAP obtained with all queries expanded. This decision method is thus safe. The behavior of Precision at 10 is more variable, depending on the choice of the threshold.

The second observation is that selective QE positively affects the NrNoRel measure. The new runs have almost the same NrNoRel performance as the unexpanded runs, and this was one of the main objectives of our investigation.

2 Web track: topic distillation task

FUB participation in the topic distillation task of the Web track focused on only-content analysis.

Last year FUB didn't participate in this task, although the same baseline Information Retrieval system was employed by the Glasgow University (GU)[8].

In particular, GU analysed both the Absorbing Model and PageRank based algorithms for link analysis on top of our baseline IR system by using different content-link score combination approaches. Using the WEB corpus .GOV some utility functions combining link and text analyses were shown to moderately improve the performance over the baseline.

However, no query expansion technique was employed by GU in these TREC 11 experiments. As we successfully used query expansion for the "topic relevance task" at TREC 10, we checked whether the same strategy was also effective for the "topic distillation task" of TREC 11. The only modification we performed was to the value of the parameter c (from $c = 7$ to $c = 1$). We found that the application of the query expansion "as it was" applied in TREC 10 was still effective to the topic distillation task of TREC 11. Indeed, as shown in Table 5 we achieved better results than those reported by the best system participating in TREC 11. For topic distillation of this year we have applied exactly the same strategy as for the experiments on the TREC 11 collection.

However, preliminary results suggest that the content-only term-weighting was not effective this year. The difference in performance is probably due to a change in the type of the task, with a different assessment of the notion of relevance. Judging from our results, the "topic distillation task of TREC 2002" looks very different from the "topic distillation task of TREC 2003".

Acknowledgments

The experiments were conducted using the first version of the Terrier's Information Retrieval platform³ initially developed by Gianni Amati during his PhD at Glasgow University.

References

- [1] Giambattista Amati. *Probability Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, Glasgow University, 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 Publication 500-250.

³<http://www.dcs.gla.ac.uk/ir/terrier/index.html>

Table 4: The selective application of QE.

Parameters	Runs with QE			
	fub03InB2e3	fub03IeOLKe3	fub03InOLe3	fub03IneOLe3
	DFR Models			
$c = 3$	I(n)B2	I(n.e)OL2	I(n)OL2	I(n.e)OL2
	DFR Expansion models			
$\beta = 0.4$	Bo2	KL	Bo2	Bo2
	all topics with QE			
@10:	0.4180	0.4070	0.4130	0.3980
MAP:	0.2434	0.2503	0.2519	0.2479
top 10 with No Rel.	18	18	17	20
topics with QE	100	100	100	100
	all topics with selective QE			
InfoQ < 0.12				
@10:	0.4230	0.3950	0.4210	0.3950
MAP:	0.2456	0.2543	0.2556	0.2524
top 10 with No Rel.	11	16	15	16
topics with QE	68	67	66	67
	all topics with selective QE			
InfoQ < 0				
@10:	0.4140	0.3950	0.4080	0.3950
MAP:	0.2439	0.2486	0.2527	0.2477
top 10 with No Rel.	11	16	14	16
topics with QE	41	41	37	41
	Baseline			
@10:	0.4080	0.3950	0.3940	0.3950
MAP:	0.2292	0.2282	0.2330	0.2282
top 10 with No Rel.	11	16	12	16
topics with QE	0	0	0	0

- [3] 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.
- [4] 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.
- [5] 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.
- [6] Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306. ACM Press, 2002.
- [7] Morris H. DeGroot. *Probability and Statistics*. Addison-Wesley, 2nd edition, 1989.
- [8] V. Plachouras, I. Ounis, G. Amati, and C.J. Van Rijsbergen. University of Glasgow at the Web track of TREC 2002. In E.M. Voorhees and L. Buckland, editors, *Proceedings of the 11th Text Retrieval Conference TREC 2002*, Gaithersburg, MD, November 2003. NIST Special Publication.
- [9] Robert G.D. Steel, Jamies H. Torrie, and David A. Dickey. *Principles and Procedures of Statistics. A Biometrical Approach*. MacGraw–Hill, 3rd edition, 1997.

Table 5: Performance of the 5 submitted runs for topic distillation task on TREC 11 and TREC 12. (QE is with 3 documents and 10 terms, $c = 1$ and $\beta = 0.5$).

Topic Distillation TREC 11			
Models	QE	MAP	Pr @10
I(n)L2	Bo1	0.2166	0.3000
I(n)L2	B	0.2194	0.3020
I(n.e)B2	B	0.2251	0.2939
I(n.e)B2	BM	0.2246	0.2939
I(n)B2	BM	0.2012	0.2918

Topic Distillation TREC 12				
Run	Models	QE Models	MAP	Pr @10
fub03InLBo1t	I(n)B2	Bo1	0.0810	0.0580
fub03InLBt	I(n)B2	B	0.0778	0.0620
fub03IneBBt	I(n.e)B2	B	0.0799	0.0640
fub03IneBMt	I(n.e)B2	BM	0.0818	0.0620
fub03InBMt	I(n)B2	BM	0.0775	0.0640