Query difficulty, robustness and selective application of query expansion

Giambattista Amati, Claudio Carpineto, and Giovanni Romano

Fondazione Ugo Bordoni
Rome Italy
gba, carpinet, romano@fub.it

Abstract. There is increasing interest in improving the robustness of IR systems, i.e. their effectiveness on difficult queries. A system is *robust* when it achieves both a high Mean Average Precision (MAP) value for the entire set of topics and a significant MAP value over its worst X topics (MAP(X)). It is a well known fact that Query Expansion (QE) increases global MAP but hurts the performance on the worst topics. A selective application of QE would thus be a natural answer to obtain a more robust retrieval system.
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 second measure is used to selectively apply QE. This method achieves a performance similar to that with unexpanded method on the worst topics, and better performance than full QE on the whole set of topics.

1 Introduction

Formulating a well-defined topic is a fundamental issue in Information Retrieval. Users in general are not aware of the intrinsic ambiguity conveyed by their queries, as well as they are not confident on whether submitting short or long queries to obtain the highest retrieval quality. It is a largely accepted evidence that, for example, pseudo-relevance feedback (also known as blind feedback or retrieval feedback) can be used to expand original queries with several additional terms with the aim of a finer formulation of the initial queries.

In many cases the QE process succeeds, but in some cases QE worsens the quality of the retrieval. Global performance values tell us the average behaviour of the system, but not if the system has a large variance in performance over all single topics. Retrieval can be excellent with some queries and very poor-performing with others. The introduction of the notion of robustness in retrieval is thus motivated by the necessity of improving the quality of the retrieval also on the most difficult queries.

Two new evaluation measures for robustness have been defined in the TREC environment, i.e. the number of topics with no relevant documents in the top retrieved 10 (denoted in this paper by NrTopicsWithNoRel) and MAP(X), which

measures the area under the average precision over the *worst* X (=%25 of) topics. A topic is deemed worst with respect to the individual run being evaluated.

The use of full QE usually results in an improvement of global MAP over the unexpanded method. However, we notice that:

- the number NrTopicsWithNoRel of topics with no relevant documents in the top retrieved 10 increases when QE is activated, and similarly
- MAP of the worst X topics diminishes when QE is adopted.

Briefly, QE always worsen the performance of the system on its poor-performing queries. The objective of our study focuses on defining a decision method for QE activation able to achieve or ameliorate as much as possible the global MAP value obtained by full QE, while keeping the other two new measures at the same value as the unexpanded method.

We address the following issues:

- defining an operational notion of a poor-performing query, that can be used to have a measure on the risk of performing QE on that query.
- defining a measure predicting on what queries there is a reasonable chance that QE fails or succeeds. This measure can be used to selectively activate QE.

The problem of predicting a poor-performing query is not new. It has been investigated under different names, such as query-difficulty, query-specificity, query-ambiguity or even as an inherent problem of QE. Indeed, the importance of the query-terms based on the quality of the first-pass ranking can be assessed. According to Kwok [**?**] the within-document term-frequency and the standard *Idf* can be combined to "peak up" those query-terms that hold a higher weight in the original query. Difficult queries are those which do not possess a variety of such important terms.

A different approach uses a similarity measure based on the cosine function to generate the query-space. This topology does not exhibit a significant regularity in the position of the difficult and of the easy queries. There is only some continuity in the position of the queries of the same type [**?**].

The closest work to our investigation is the clarity score based on the language model [**?**]. The clarity score needs to be computed at the indexing time since it contains a collection model component.

As far as we know there has not been any methodological or experimental work addressing the problem of the selective application of QE. This is a challenging task because it requires to formally relate the relevance to other notions like query-difficulty and query expansion.

We propose a method which achieves a performance similar to that of the unexpanded baseline on the worst topics, but better performance than full QE

on the whole set of topics. Our work is thus a fist step towards the definition of a decision methodology for the selective use of the QE.

In our analysis we use the DFR (*Divergence From Randomness*) modular probabilistic framework [**?**,**?**,**?**] together with query expansion based on distribution analysis [**?**,**?**,**?**] for retrieval. We use the data of the TREC 2003 robust track.

2 Retrieval Setting

Our objective is to develop a methodology for a stable and robust QE activation to improve the performance on both the worst topics and all topics. We consider the description-only queries of the TREC 2003 robust track data. Their average length after stemming and with the stop list is about 8 terms.

The retrieval framework is made up of two components: the DFR within-document term-weighting model and QE within-query term-weighting model. In the next two sections we briefly describe these two components. We use different DFR models to test robustness with selective QE activation.

2.1 Term-weighting models

The DFR within-document term-weighting models are:

I(n)OL2, I(n_e)OL2, I(n)B2, I(n_e)B2, I(n_e)OB2. They are obtained from the generating formula:

$$\text{Info}_{\text{DFR}} = -\log_2 \text{Prob}(\text{term\_freq}|\text{doc\_freq}, \text{Freq}(\text{term}|\text{Collection})) \quad (1)$$

where Prob is the probability of obtaining a given within-document term-frequency randomly. Formula 1 is not used directly, but it is normalized by considering the probability of the observed term-frequency only in the set of documents containing the term. The final weighting formulas are:

$$\text{I(n)OL2}: \quad \frac{tfn}{tfn+1}\log_2\left(\frac{N-\text{doc\_freq}+1}{\text{doc\_freq}+0.5}\right) \quad (2)$$

$$\text{I(n}_\text{e}\text{)OL2}: \quad \frac{tfn}{tfn+1}\log_2\left(\frac{N-n_e+1}{n_e+0.5}\right) \quad (3)$$

$$\text{I(n)B2}: \quad \frac{\text{Freq}(\text{term}|\text{Collection})+1}{\text{doc\_freq}\cdot(tfn+1)}\left(tfn\cdot\log_2\frac{N+1}{\text{doc\_freq}+0.5}\right) \quad (4)$$

$$\text{I(n}_\text{e}\text{)B2}: \quad \frac{\text{Freq}(\text{term}|\text{Collection})+1}{\text{doc\_freq}\cdot(tfn+1)}\left(tfn\cdot\log_2\frac{N+1}{n_e+0.5}\right) \quad (5)$$

$$\text{I(n}_\text{e}\text{)OB2}: \quad \frac{\text{Freq}(\text{term}|\text{Collection})+1}{\text{doc\_freq}\cdot(tfn+1)}\left(tfn\cdot\log_2\frac{N-n_e+1}{n_e+0.5}\right) \quad (6)$$

where

$tfn = \text{term\_freq} \cdot \log_2 \left(1 + c \cdot \frac{\text{average\_document\_length}}{\text{document\_length}}\right)$,
N is the size of the collection,
$n_e = N \cdot \left(1 - \left(\frac{1}{N}\right)^{\text{Freq(term|Collection)}}\right)$,
Freq(term|Collection) is the within-collection term-frequency,
term_freq is the within-document term-frequency,
doc_freq is the document-frequency of the term,
the parameter c is set to 3.

2.2 Query expansion

The QE method is the same as used an TREC-10 with very good results[**?**] 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. $\frac{tfq}{max_{t \in q} tfq}$
$\text{MaxInfo} = \arg_{t \in q^*} \max \text{Info}_{\text{DFR}}$
Info_{DFR} is a term-frequency in the expanded query induced by using a DFR model, that is:

$$\text{Info}_{\text{DFR}} = -\log_2 \text{Prob}(\text{Freq}(\text{term}|\text{TopDocuments})|\text{Freq}(\text{term}|\text{Collection})) \quad (7)$$

Formula 7 uses the same probabilistic model Prob of Formula 1, but the observed frequencies are different. The term-weighting models compute the probability of obtaining a given within-document term-frequency, whereas the within-query term-weighting computes the probability of obtaining a given term-frequency within the topmost retrieved documents.

For the implementation of Info_{DFR} we here use the normalized Kullback-Leibler measure (KL) [**?**,**?**]

$$\text{Info}_{\text{KL}}(\text{t}) = \frac{\text{Freq(t|TopDocs)}}{\text{TotFreq(TopDocs)}} \cdot \log_2 \frac{\text{Freq(t|TopDocs)} \cdot \text{TotFreq(C)}}{\text{TotFreq(TopDocs)} \cdot \text{Freq(t|C)}} \quad (8)$$

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

Parameters	Models with full QE				Model without QE
$c = 3$	I(n)B2	I(n_e)OL2	I(n)OL2	I(n_e)OL2	I(n_e)OB2
	DFR Expansion models				
$\beta = 0.4$	Bo2	KL	Bo2	Bo2	-
	100 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

where C indicates the whole collection and TopDocs denotes the pseudo-relevant set, while the Bose-Einstein statistics (Bo2) is:

$$\text{Info}_{\text{Bo2}}(\text{t}) = -\log_2\left(\tfrac{1}{1+\lambda}\right) - \text{Freq}(\text{t}|\text{TopDocuments}) \cdot \log_2\left(\tfrac{\lambda}{1+\lambda}\right) \quad \text{[Bo2]}$$

$$\lambda = \text{TotFreq}(\text{TopDocuments}) \cdot \frac{\text{Freq}(\text{t}|\text{Collection})}{\text{TotFreq}(\text{Collection})} \qquad (9)$$

A further condition imposed for the selection of the new query-terms is that they must appear in at least two retrieved documents. This condition is to avoid the noise that could be produced by those highly informative terms which appear only once in the set of the topmost retrieved documents. The QE parameters are set as follows:

$\beta = 0.4$

$|\text{TopDocuments}| = 10$

the number of terms of the expanded query is equal to 40.

Table 1 compares a baseline run with the full QE runs. We chose the model I(n_e)OB2 defined in Formula 6 as baseline for the comparison, since it is the " best" performing model on the most difficult topics.

The unexpanded runs achieve the best MAP(X) and the lowest NrTopicsWithNoRel, and the runs with expanded queries achieve the highest values of MAP and precision at 10.

3 Selective application of QE

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

We exploit the Info_{DFR} measures, as defined by Formula 7, and introduce a new measure InfoQ. We show that the sum of all Info_{DFR} over the terms of the query is related to the Average Precision (AP) and InfoQ is related to the

AP increase after the QE activation. In other words, Info $_{\text{DFR}}$ is an indicator of a possible low outcome of AP, attesting thus when a topic is possibly *difficult*. On the other hand, InfoQ is an indicator of the successful application of QE.

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

3.1 Test data

The document collection used to test robustness is the set of documents on both TREC Disks 4 and 5 minus the Congressional Record on disk 4, containing 528,155 documents of 1.9 GB size. The set of test-topics contains 100 statements. Among these topics there are 50 topics that are known to be difficult for many systems. These 50 difficult topics were extracted from all 150 queries of previous TRECs using this same collection. We have indexed all fields of the documents and used Porter's stemming algorithm.

3.2 How QE affects Robustness

Consider as an example the performance of the model of Formula 2, I(n)OL2, as shown in Table 2.

With full QE, we achieve 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 MAP 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 $\sim$0.063 and the gain is much greater than the average loss ($\sim$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.

Comparing the figures of Table 2 with those relative to all the 150 queries of the past TREC data, we observe a detriment of the success rate. The success rate is around 65% with all the 150 old queries of past TREC data. A detriment in precision at 10 is 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 is larger ($\sim$ +10%) than that obtained with this TREC data ($\sim$ +5%).

3.3 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 top-most documents in the ranking. If the early precision of the first-pass retrieval is

Table 2. Run I(n)OL2 with description-only topics. The columns with "No QE" contain the number of queries to which the QE was not applied.

100 Topics										
Baseline		I(n)OL2 with QE				I(n)OL2 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

high, then we have a good chance to extract good additional topic terms together with their relative query-weights. To start our investigation we first compute 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 [**?**]. A negative correlation factor indicates that the two variables are inversely related.

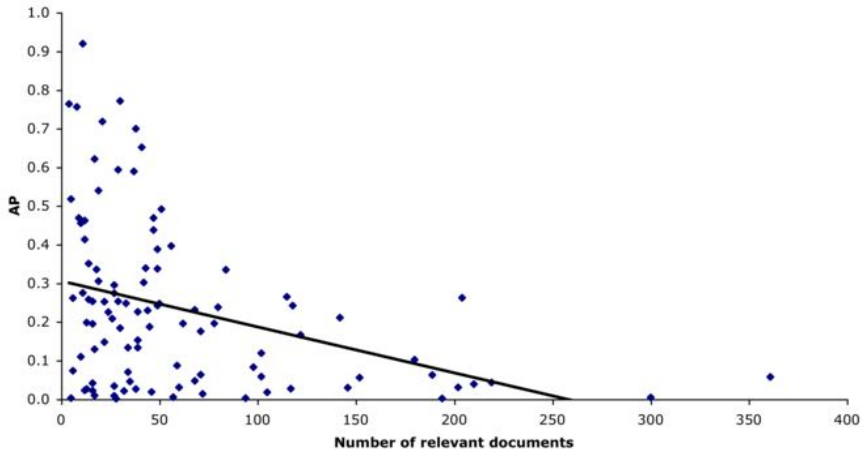
Surprisingly, these correlation factors come out to be both negative:
$\rho(Rel, \text{AP}) = -0.36$ and $\rho(Rel, \text{P@10}) = -0.14$.

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 [**?**].

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.

Fig. 1. The number of relevant documents is inversely related to AP of the unexpanded query ($\rho(Rel, \text{AP}) = -0.36$). 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.

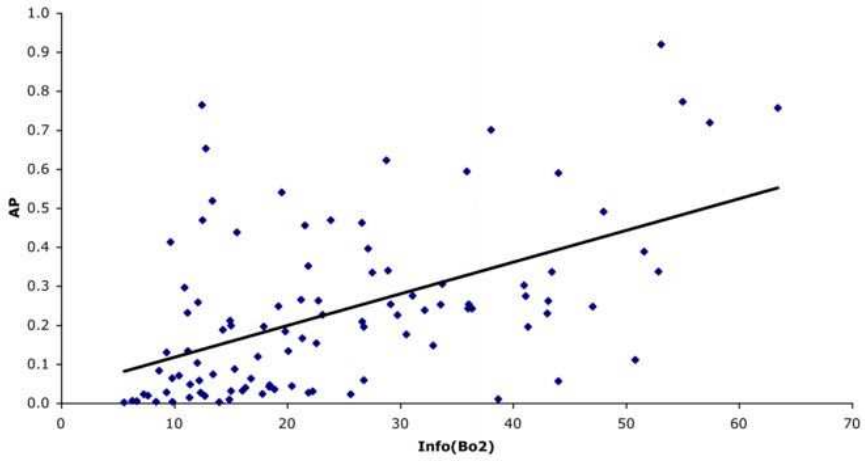
$$\text{Info}_{\text{DFR}} = \sum_{\text{t} \in Q} - \log_2 \text{Prob}(\text{Freq}(\text{t}|\text{TopDocuments})|\text{Freq}(\text{t}|\text{Collection})) \quad (10)$$

where $\text{Prob}(\text{t}|\text{Collection}, \text{TopDocuments})$ is a DFR basic model (based on the Binomial, the Bose-Einstein statistics or the Kullback-Leibler divergence measure, as defined in Formulas 8 and 9). We here use the probability of Bose-Einstein defined in Formula (9). We stress agin the fact that the same weighting formula is used by our expansion method. together with the Kullback-Leibler divergence I(n_e)OL2 (see Table 1).

There are other information theoretic measures capturing the notion of term–specificity of the query.

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

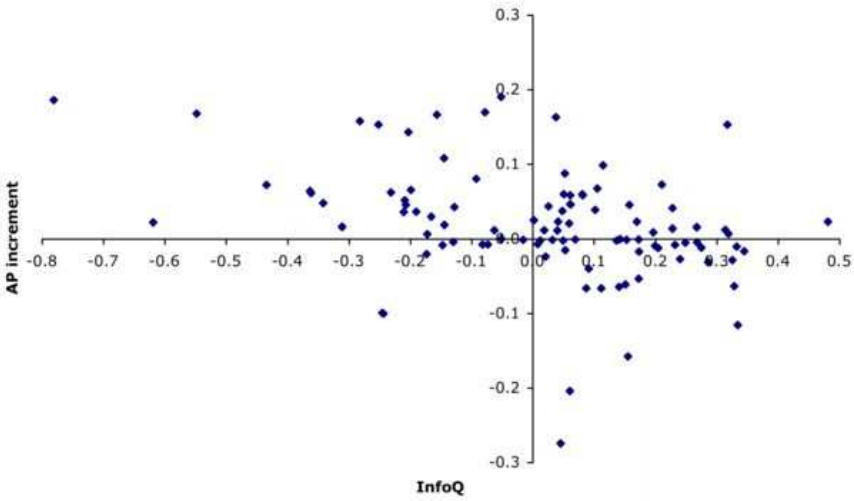
Fig. 2. The information content Info $_{\text{Bo2}}$ of the query within the topmost retrieved documents is linearly correlated to the AP of the unexpanded queries ($\rho(\text{Info}_{\text{Bo2}}, \text{AP}) = 0.52$). Specific queries have a large value of Info $_{\text{DFR}}$.



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 $_{\text{DFR}}$ has a significant correlation with the AP of the unexpanded queries $\rho(\text{Info}_{\text{Bo2}}, \text{AP}) = 0.52$. Similarly to the negative correlation between the number of relevant documents and the AP of the unexpanded queries, which is $\rho(Rel, \text{AP}) = -0.36$, the correlation factor between the score InfoQ and Rel is negative ($\rho(Rel, \text{Info}_{\text{Bo2}}) = -0.23$). Again, this may be explained by the fact that specific queries possess fewer relevant documents.

Unfortunately, we did not find a significant correlation between Info $_{\text{DFR}}$ and QE; that is, Info $_{\text{DFR}}$ 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 [**?**] that although the retrieval effectiveness of QE in general increases as the query difficult decreases, very easy queries hurt performance.

Fig. 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 ($\rho(\text{QE increase rate, InfoQ}) = -0.33$). The first and the third quadrants contain the errors when the threshold is set to 0.



3.4 Predicting the successful application of QE

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

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

is shown to have a moderately weak negative correlation with QE:

$\rho(\text{QE, InfoPriorQ}) = -0.27$.

InfoPriorQ is linearly related to the length of the query with correlation factor $\rho(\text{QueryLength, InfoPriorQ}) = 0.90$, 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

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

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

The function:

$$\text{InfoQ} = \frac{1}{\text{QueryLength}} \left(\frac{\text{InfoPriorQ} - \mu_{\text{InfoPriorQ}}}{\sigma_{\text{InfoPriorQ}}} + M_Q \right) \tag{11}$$

where the μ_Xs and the σ_Xs stand for the mean and the standard deviation of the X values, combines InfoPriorQ and Info $_{\text{DFR}}$. Info $_{\text{DFR}}$ query rankings may not agree using different DFR models. Because the correlation factor is negative, and since we trigger the QE when InfoQ is below a given threshold, a cautious way to smooth different Info $_{\text{DFR}}$ 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 length[1] ($\rho(\text{QE},\ \text{InfoQ}) = 0.62$).

4 Discussion of results

In Table 4 we summarize the results on the selective application of QE. The MAP(X) values are not reported since the new values are similar to those in the full QE models; thus we focus on the other measures. We compare the performance of models with full QE with the performance of the models with selective QE under the same setting.

The first remark is that the decision rule for QE activation is quite robust. The MAP of models with selective QE is greater than the MAP of the full QE models for a large range of values of the threshold 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

[1] Using $\log_2(\text{QueryLength})$ instead of QueryLength the score of Formula 11 is more correlated to the query length with $\rho(\text{QueryLength},\ \text{InfoQ}) = 0.74$ and $\rho(\text{QE},\ \text{InfoQ}) = -0.34$.

Table 4. The selective application of QE.

Parameters	Runs with QE			
	I(n)B2	I(n_e)OL2	I(n)OL2	I(n_e)OL2
	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
InfoQ < 0.12	all topics with selective QE			
@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
InfoQ < 0	all topics with selective QE			
@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

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 NrTopicsWithNoRel measure. The models with selective QE have almost the same NrTopicsWithNoRel performance as the unexpanded runs, and this is one of the main objectives of our investigation.

5 Conclusions

We have defined two information theoretic functions used to predict the query-difficulty and to selectively apply QE. Our objective was to avoid the application of QE on the set of worst (difficult) topics. Indeed, QE application predictor achieves a performance similar to that of the unexpanded method on the worst topics, and better performance than full QE on the whole set of topics. Our work is thus a promising step towards a decision methodology for the selective use of the QE.

Acknowledgments

The experiments were conducted using the first version of the Terrier's Information Retrieval platform (http://www.dcs.gla.ac.uk/ir/terrier) initially developed by Gianni Amati during his PhD at Glasgow University. Terrier is a modular Information Retrieval framework that provides indexing and retrieval functionalities.

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 Pubblication 500-250.
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. K. L. Kwok. A new method of weighting query terms for ad-hoc retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 187–195. ACM Press, 1996.
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.
10. Terry Sullivan. Locating question difficulty through explorations in question space. In *Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries*, pages 251–252. ACM Press, 2001.