Comparing Weighting Models for Monolingual Information Retrieval

Gianni Amati, Claudio Carpineto, and Giovanni Romano

Fondazione Ugo Bordoni, via B. Castiglione 59,
00142 Rome, Italy
{gba,carpinet,romano}@fub.it

Abstract. Motivated by the hypothesis that the retrieval performance of a weighting model is independent of the language in which queries and collection are expressed, we compared the retrieval performance of three weighting models, i.e., Okapi, statistical language modeling (SLM), and deviation from randomness (DFR), on three monolingual test collections, i.e., French, Italian, and Spanish. The DFR model was found to consistently achieve better results than both Okapi and SLM, whose performance was comparable. We also evaluated whether the use of retrieval feedback improved retrieval performance; retrieval feedback was beneficial for DFR and Okapi and detrimental for SLM. Besides relative performance, DFR with retrieval feedback achieved excellent absolute results: best run for Italian and Spanish, third run for French.

1 Introduction

Although the choice of the weighting model may crucially affect the performance of any information retrieval system, there has been little work on evaluating the relative merits and drawbacks of different weighting models in the CLEF environment. The main goal of our participation in CLEF03 was to help fill this gap.

We consider three weighting models with a different theoretical background that have proved their effectiveness on a number of tasks and collections. The three models are Okapi [9], statistical language modeling [11], and deviation from randomness [1].

We study the retrieval performance of the rankings produced by each weighting model with and without retrieval feedback, on three monolingual test collections, i.e., French, Italian and Spanish. The collections are indexed with standard techniques and the retrieval feedback stage is performed using the method described in [5].

In the following we first describe the three weighting models, the method used for retrieval feedback, and the experimental setting. Then we compare the retrieval performance of the three methods, performing also a query-by-query analysis. Finally, we summarize the main results of the experiments.

2 The three weighting models

For the ease of clarity and comparison, the document ranking produced by each weighting model is represented using the same general expression, namely as the product of a document-based term weight by a query-based term weight:

$$sim(q,d) = \sum_{t \in q \wedge d} w_{t,d} \cdot w_{t,q}$$

This formalism also allows a uniform application of the subsequent retrieval feedback stage to the first-pass ranking produced by each weighting model, as we will see in the next section. Before giving the expressions for $w_{t,d}$ and $w_{t,q}$ for each weighting model, we report the complete list of variables that will be used:

f_t	the number of occurrences of term t in the collection
$f_{t,d}$	the number of occurrences of term t in document d
$f_{t,q}$	the number of occurrences of term t in query q
n_t	the number of documents in which term t occurs
D	the number of documents in the collection
T	the number of terms in the collection
λ_t	the ratio between f_t and T
l_d	the length of document d
l_q	the length of query q
avr_l_d	the average length of documents in the collection

2.1 Okapi

To describe Okapi, we use the expression given in [9]. This formula has been used by most participants in TREC and CLEF over the last years.

$$w_{t,d} = \frac{(k_1+1) \cdot f_{t,d}}{k_1 \cdot \left[(1-b) + b \, \frac{l_d}{avr\_l_d}\right] + f_{t,d}}$$

$$w_{t,q} = \frac{(k_3+1) \cdot f_{t,q}}{k_3 + f_{t,q}} \cdot log_2 \frac{D - n_t + 0.5}{n_t + 0.5}$$

2.2 Statistical Language Modeling (SLM)

The statistical language modeling approach has been proposed in several papers, with many variants (e.g., [6], [7]). Here we use the expression given in [11], with Dirichlet smoothing.

$$w_{t,d} = log_2 \frac{f_{t,d} + \mu\lambda_t}{l_d + \mu} - log_2 \frac{\mu}{l_d + \mu} - log_2 \lambda_t + \frac{l_q}{|q \wedge d|} \cdot log_2 \frac{\mu}{l_d + \mu}$$

$$w_{t,q} = f_{t,q}$$

2.3 Deviation From Randomness (DFR)

Deviation from randomness has been successfully used at CLEF 2002, for the Italian monolingual task [1], and at TREC, for the Web and Robust tracks ([2], [3]). It is best described in [4].

$$w_{t,d} = (log_2(1 + \lambda_t) + f^*_{t,d} \cdot log_2\frac{1+\lambda_t}{\lambda_t}) \cdot \frac{f_t + 1}{n_t \cdot (f^*_{t,d} + 1)}$$

with

$$f^*_{t,d} = f_t \cdot log_2(1 + \frac{c \cdot avr\_l_d}{l_d})$$

3 Retrieval feedback

As retrieval feedback has been incorporated in most recent systems participating in CLEF, it is interesting to also evaluate the performance of the different weighting models when they are enriched with retrieval feedback.

To perform the experiments, we used information-theoretic query expansion [5]. At the end of the first-pass ranking, each term in the top retrieved documents was assigned a score using the Kullback-Leibler distance between the distribution of the term in such documents and the distribution of the same term in the entire collection, and the terms with the highest scores were selected for expansion. The KLD scores are given by:

$$KLD_{t,d} = f_{t,d} \cdot log_2\frac{f_{t,d}}{f_t}$$

At this point, the KLD scores were also used to reweight the terms in the expanded query. As the weights for the unexpanded query (i.e., SLM, Okapi, and DFR) and the KLD scores had different scales, we normalized both the weights of the original query and the scores of the expansion terms by the maximum corresponding value; then the normalized values were linearly combined. The new expression for computing the similarity between an expanded query q_{exp} and a document d becomes:

$$sim(q_{exp}, d) = \sum_{t \in q \wedge d} w_{t,d} \cdot (\alpha \frac{w_{t,q}}{Max_q w_{t,q}} + \beta \frac{KLD_{t,d}}{Max_d KLD_{t,d}})$$

4 Experimental Setting

4.1 Test Collections

The experiments were performed using three CLEF 2003 monolingual test collections, namely the French, Spanish, and Italian collections. For all collections, the title+description topic statement was considered.

4.2 Document and Query Indexing

We identifed the individual words occurring in the documents, considering only the admissible sections and ignoring punctuation and case. The system then performed word stemming and word stopping. For word stemming, we used the French, Italian, and Spanish versions of Porter stemming algorithm [8], which have been made available on the Snowball web site (http://snowball.tartarus.org) To remove common words, we used the stop lists provided by Savoy [10]. Thus, we performed a strict single-word indexing; furthermore, we did not use any ad hoc linguistic manipulation such as expanding or removing certain words from the query text or using lists of proper nouns.

4.3 Choice of Experimental Parameters

The final document ranking is affected by a number of parameters. To perform the experiments, we set the parameters using values that have been reported in the literature. Here is the complete list of parameter values:

Okapi	$k_1 = 1.2$, $k_3 = 1000$, $b = 0.75$
SLM	$\mu = 1000$
DFR	$c = 2$
Retrieval feedback	10 pseudo-rel. docs., 40 exp. terms, $\alpha = 1$, $\beta = 0.5$

5 Results

For each collection and for each query, we computed six runs: two runs for each of the three weighting modesl, one without and one with retrieval feedback (RF). Table 1, Table 2, and Table 3 show the retrieval performance of each method on the French, Italian, and Spanish collection, respectively. Performance was measured using average precision (AV-PREC), precision at 5 retrieved documents (PREC-AT-5), and precision at 10 retrieved documents (PREC-AT-10). For each collection we show in bold the best result with retrieval feedback and the best result without retrieval feedback.

Table 1. Retrieval performance on the French collection

	AV-PREC	PREC-AT-5	PREC-AT-10
Okapi	0.5030	0.4385	**0.3654**
Okapi + RF	0.5054	0.4769	0.3942
SLM	0.4753	0.4538	0.3635
SLM + RF	0.4372	0.4192	0.3462
DFR	**0.5116**	**0.4577**	**0.3654**
DFR + RF	**0.5238**	**0.4885**	**0.3981**

Table 2. Retrieval performance on the Italian collection

	AV-PREC	PREC-AT-5	PREC-AT-10
Okapi	0.4762	0.4588	0.3510
Okapi + RF	0.5238	0.4824	0.3902
SLM	0.5027	**0.4941**	**0.3824**
SLM + RF	0.5095	0.4824	0.3863
DFR	**0.5046**	0.4824	0.3725
DFR + RF	**0.5364**	**0.5255**	**0.4137**

Table 3. Retrieval performance on the Spanish collection

	AV-PREC	PREC-AT-5	PREC-AT-10
Okapi	0.4606	0.5684	0.5175
Okapi + RF	0.5093	0.6105	0.5491
SLM	0.4720	**0.6140**	0.5157
SLM + RF	0.5112	0.5825	0.5316
DFR	**0.4907**	0.6035	**0.5386**
DFR + RF	**0.5510**	**0.6140**	**0.5825**

Note that for the French and Italian collections the average precision was greater than the early precisions; this is due to the fact that for these collections the mean number of relevant documents per query is, on average, small, and that there are many queries with very few relevant documents.

The first main finding of our experiments is that the best absolute result for each collection and for each evaluation measure was always obtained by DFR with retrieval feedback, with notable improvements on several data points. The excellent performance of the DFR model is confirmed when comparing the weighting models without query expansion, although in the latter case DFR did not always achieve the best results (i.e., for PREC-AT-5 and PREC-AT-10 on Italian, and for PREC-AT-5 on Spanish).

Of the other two models (i.e., Okapi and SLM), none was clearly superior to the other. They achieved comparable results on Spanish, while Okapi was slightly better than DFR on French and slightly worse on Italian. However, when considering the first retrieved documents, the performance of SLM was usually very good and sometimes even better than DFR.

The results in Table 1, Table 2, and Table 3 show also that retrieval feedback improved Okapi and DFR runs and mostly hurt SLM runs. In particular, the use of retrieval feedback improved the retrieval performance of Okapi and DFR for all evaluation measures and across all collections, whereas it usually decreased the early precision of SLM and on one occasion (i.e., for French) it hurt even the average precision of SLM. The unsatisfying performance of SLM + RF may be explained by considering that the experiments were performed using long queries.

We would like to emphasize that the DFR runs shown here correspond to actually submitted runs, although they were not our best runs. In fact, our best submitted runs had language-specific optimal parameters tuned using the past CLEF collections. Then we submitted for each language a run with the same experimental parameters, obtained by averaging the best parameters.

The parameters of our best runs were as follow. For French, $c = 2$, number of pseudo-relevant documents = 8 , number of expansion terms = 30, $\alpha = 1$, $\beta = 0.25$; run *fub03fr3*, average precision = 0.5377, ranked as the third absolute run. For Italian, $c = 2$, number of pseudo-relevant documents = 10 , number of expansion terms = 40, $\alpha = 1$, $\beta = 0.5$; run *fub03itB*, average precision = 0.5707, ranked as the first absolute run. For Spanish, $c = 2$, number of pseudo-relevant documents = 5 , number of expansion terms = 50, $\alpha = 1$, $\beta = 0.5$; run *fub03itB*, average precision = 0.5533, ranked as the first absolute run.

We also performed a query-by-query analysis. For each query, we computed the difference between the best and the worst retrieval result, considering average precision as the performance measure. Figure 1, Figure 2, and Figure 3 show the results for French, Italian, and Spanish, respectively.

Thus, the length of each bar depicts the range of performance variations attainable by the three methods (with retrieval feedback) for each query. The results show that the intermethod variations on sigle queries was ample, but does not tell us which method performed best.

To get a more complete picture, we counted, for each collection, the number of queries for which each method achieved the best, median, or worst performance. The results, shown in Table 4, confirm the better retrieval e ectiveness of DFR over the other two models. The superiority of DFR over Okapi and SLM was clear for Spanish, while DFR and Okapi obtained more comparable results on the other two test collections. For French and Italian, the number of best results obtained by DFR and Okapi was similar, but, on the whole, DFR was ranked ahead of Okapi for a much larger number of queries.

Table 4. Ranked performance

	French			*Italian*			*Spanish*		
	SLM	Okapi	DFR	SLM	Okapi	DFR	SLM	Okapi	DFR
1st	11	20	21	10	21	20	16	16	25
2nd	11	17	24	9	16	26	10	22	25
3rd	30	15	7	32	14	5	31	19	7

6 Results

The main conclusion of our experiments is that the DFR model was more effective than both Okapi and SLM, which achieved comparable retrieval performance. In particular, DFR with query expansion obtained the best average absolute results for any evaluation measure and across all test collections.

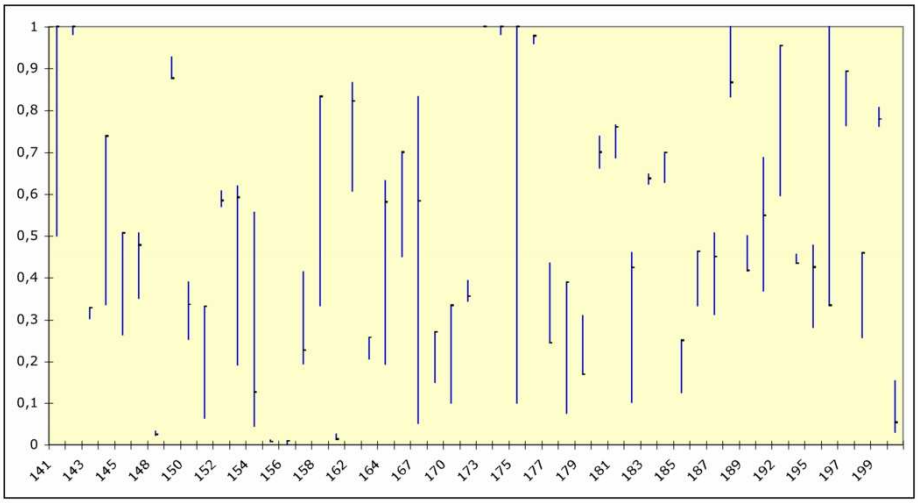

Fig. 1. Performance variation on individual queries for French

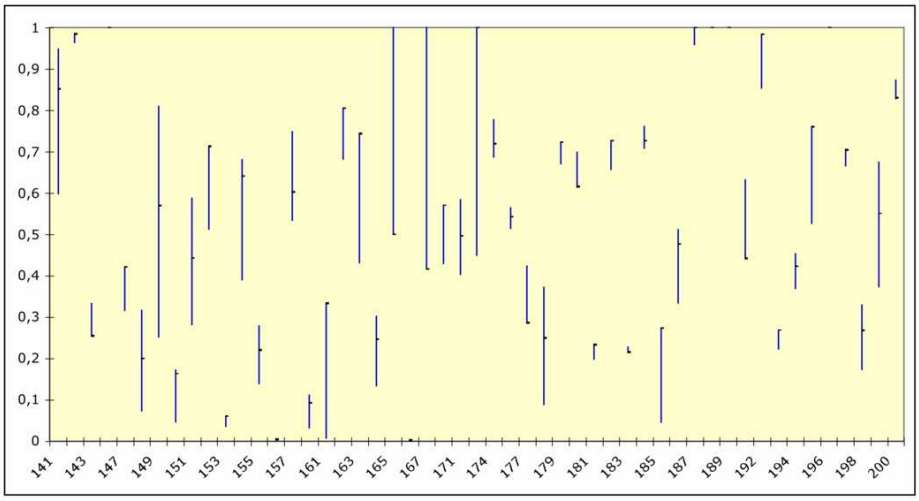

Fig. 2. Performance variation on individual queries for Italian

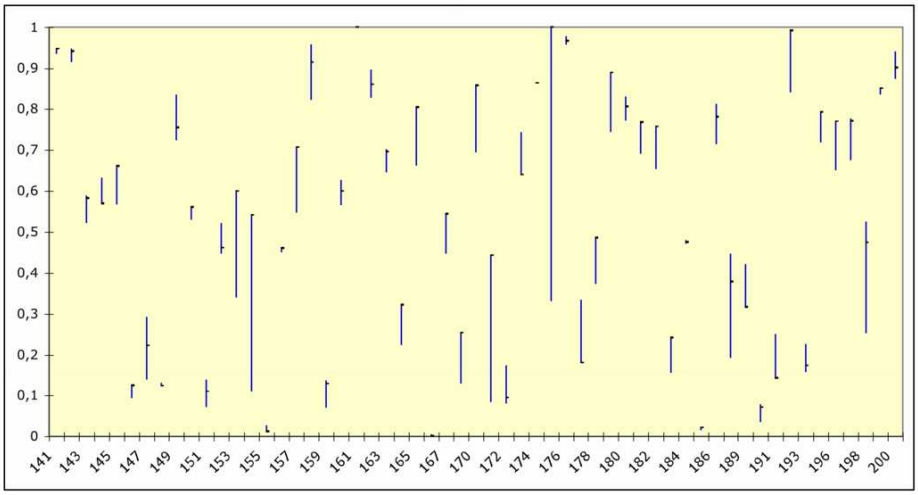

Fig. 3. Performance variation on individual queries for Spanish

The second conclusion is that retrieval feedback always improved the performance of Okapi and DFR, whereas it was often detrimental to the retrieval effectiveness of SLM, although the latter finding may have been influenced by the length of the queries used in the experiments.

These results seem to suggest that the retrieval performance of a weighting model is only moderately affected by the choice of the language, but this hypothesis should be taken with caution, because our results were obtained under specific experimental conditions.

Although there are reasons to believe that similar results might hold also across different experimental situations, in that we chose simple and untuned parameter values and made typical indexing assumptions, the issue needs more investigation. The next step of this research is to experiment with a wider range of factors, such as the length of queries, the values of each weighting model's parameters, and the combination of parameter values for retrieval feedback. It would also be useful to experiment with other languages, to see if the hypothesis that the retrieval performance of a weighting model is independent of the language receives further support.

References

1. G. Amati, C. Carpineto, and G. Romano. Italian monolingual information retrieval with prosit. In *Proceedings of CLEF (Cross Language Evaluation Forum 2002*, pages 182–191, Rome, Italy, 1992.
2. G. Amati, C. Carpineto, and G. Romano. Fub at trec-10 web track: A probabilistic framework for topic relevance term weighting. In *Proceedings of the 10th Text REtrieval Conference (TREC-10), NIST Special Publication 500-250*, pages 182–191, Gaithersburg, MD, USA, 2001.

3. G. Amati, C. Carpineto, and G. Romano. Fub at trec-12: Robust and web track. In *Proceedings of the 12th Text REtrieval Conference (TREC-10)*, Gaithersburg, MD, USA, 2003.
4. G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems*, 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. D. Hiemstra and W. Kraaij. Twenty-one at trec-7: Ad hoc and cross-language track. In *Proceedings of the 7th Text REtrieval Conference (TREC-7), NIST Special Publication 500-242*, pages 227–238, Gaithersburg, MD, USA, 1998.
7. J. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Reasearch and Development in Information Retrieval*, pages 275–281, 1998.
8. M. F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.
9. S. E. Robertson, S. Walker, and M. M. Beaulieu. Okapi at trec-7: Automatic ad hoc, filtering, vlc, and interactive track. In *Proceedings of the 7th Text REtrieval Conference (TREC-7), NIST Special Publication 500-242*, pages 253–264, Gaithersburg, MD, USA, 1998.
10. J. Savoy. Reports on clef-2001 experiments. In *Working Notes of CLEF 2001*, Darmstadt, Germany, 2001.
11. C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342, New Orleans, LA, USA, 2001.