Italian monolingual information retrieval with PROSIT

Gianni Amati[1], Claudio Carpineto[1], and Giovanni Romano[1]

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

Abstract. PROSIT (PRObabilistic Sifting of Information Terms) is a novel probabilistic information retrieval system that combines a term-weighting model based on deviation from randomness with information-theoretic query expansion. PROSIT has demonstrated to be highly effective at retrieving web documents at the recent TREC-10 and a version of the system is currently available as the search engine of the web site of the Italian Ministry of Communications (http://www.comunicazioni.it). In this paper, we report on the application of PROSIT to the Italian monolingual task at CLEF. We experimented with both standard PROSIT and with some enhanced versions of it. In particular, we investigated the use of bigrams and coordination level-based retrieval within the PROSIT framework. The main finding of our research are that (i) standard PROSIT was quite effective, with an average precision of 0.5116 on CLEF 2001 queries and 0.5019 on CLEF 2002 queries, (ii) bigrams were useful provided that they were incorporated into the main algorithm, and (iii) the benefits of coordination level-based retrieval were unclear.

1 Introduction

Recent research has shown the effectiveness of deviation from randomness [2] and information-theoretic query expansion [6]. We combined these techniques into a comprehensive document ranking system named PROSIT, which stands for PRObabilistic Sifting of Information Terms.

The best features of PROSIT are that it is fast, it can be easily understood and replicated, it does not virtually require training or parameter tuning, it does not employ any ad hoc linguistic manipulations, and, perhaps even more important, it has shown to be surprisingly effective.

An earlier version of PROSIT was tested at the web track of TREC-10 [1], where it was ranked as the first system for the topic relevance task. Subsequently, we developed a web version of PROSIT to act as the search engine of the web site of the Italian Ministry of Communications (http://www.comunicazioni.it). The search engine has been running on the site since the end of July 2002.

The study reported in this paper is a follow-up on this previous research, aiming at extending the scope of applications of PROSIT and improving its performance. As the site search engine based on PROSIT receives queries typically

expressed in Italian and retrieves relevant documents mostly from Italian web pages, we were particularly interested in evaluating the effectiveness of PROSIT for the Italian language. This was the first goal of our participation in CLEF.

In addition, as PROSIT performs single-word indexing and does not take advantage of the structure of queries and documents, we intended to investigate the use of multi-word indexing and text structure to improve its performance. This was our second main goal.

In the rest of the paper, we first describe the indexing stage and the two main components of PROSIT. Then we discuss how to enhance PROSIT with bigrams and coordination level-based retrieval. Finally, we present the performance results and draw some conclusions.

2 Indexing

Our system first identified 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.

Similar to earlier results reported at CLEF 2001, we found that stemming improves performance. We used a simple form of stemming for conflating singular and plural words to the same root [8]. It is likely that the use of a more sophisticated stemmer such as the Italian version of Porter's would produce better results, but we did not have a chance to try it. To remove common words, we used the stop list provided by Savoy [8].

We did not use any ad hoc linguistic manipulation such as removing certain words from the query text, e.g., "trova" (find), "documenti" (documents), etc., or expanding acronyms, e.g., does AI stand for Amnesty international or Artificial intelligence?, or using lists of proper nouns, e.g, Alberto Tomba.

The use of such manipulations makes it difficult to evaluate the overall results and makes it even more difficult to replicate the experiments. We think that it would be better to discourage their use unless it is supported by some linguistic resources which are public or which are made available by the authors.

3 Description of PROSIT

PROSIT consists of two main components: the retrieval-matching function module and the automatic query expansion module. The system has been implemented in ESL, a Lisp-like language that is automatically translated into ANSI C and then compiled by gcc compiler. PROSIT is able to index two gigabytes of documents per hour and allows sub-seconds searches on a 550 MHz Pentium III with 256 megabytes of RAM running Linux. PROSIT has been released as an application for searching documents on WEB collection. In the following two sections we explain in details the two main components of the system.

3.1 Term weighting

PROSIT can implement different matching functions which have similar and excellent performance. These basic retrieval functions are generated from a unique probabilistic framework [2]. The appealing feature of the basic term weighting models is the absence of parameters which should be learned and tuned by means of a relevance training set. In addition, the framework is easy to implement since the models use up a total number of 6 random variables which are given by the collection statistics, namely:

tf	the within document term frequency
N	the size of the collection
n_t	the size of the elite set E_t of the term (see below)
F_t	the total number of term occurrences in its elite set
l	the length of the document
avg_l	the average length of documents

However, for both TREC-10 and CLEF collections we have introduced a parameter for document length normalization which enhances the retrieval outcome.

The framework is based on computing the information gain for each term query. The information gain is obtained by a combination of three distinct probabilistic processes: the probabilistic process computing the amount of the information content of the term with respect to the entire collection, the probabilistic process computing a conditional probability of occurrence of the term with respect to an "Elite" set of documents, which is the set of documents containg the query term, and the probabilistic process deriving the term frequency within the document normalized to the document average length. The framework thus consists of three independent components: the "information content" component relative to the entire data collection, the "information gain" normalization factor component relative to the elite set of the observed term, and the "term frequency normalization function" component relative to the document length.

Our formulation of informative gain is:

$$w = (1 - Prob_1) \cdot (-log_2 Prob_2) \tag{1}$$

In our experiments we have instantiated our framework choosing Laplace's law of succession for $Prob_1$ as the gain normalization factor and the Bose-Einstein satistics for $Prob_2$:

$$Prob_1 = \frac{tfn}{tfn + 1} \tag{2}$$

$$Prob_2 = \left(\frac{1}{1+\lambda}\right) \cdot \left(\frac{\lambda}{1+\lambda}\right)^{tfn} \tag{3}$$

where:

λ is the term frequency in the collection $\frac{F_t}{N}$

tfn is the normalized term frequency $tf \cdot \log_2 \left(1 + \frac{avg\_l}{l}\right)$.

The correcting factor $c = 3$ may be inserted to the term frequency normalization and obtain $tfn = tf \cdot \log_2 \left(1 + \frac{c \cdot avg\_l}{l}\right)$. The system displayed in Formulas 1, 2 and 3 was used in our experiments and is called $B_E L2$ (B_E stands for Bose-Einstein and L for Laplace).

3.2 Retrieval feedback

Formulas 1, 2 and 3 produce a first ranking of possible relevant documents. The topmost ones are candidates to be assessed relevant and therefore we might consider them to constitute a second different "Elite set T of documents", namely documents which best describe the content of the query. We have considered in our experiments only 10 documents as pseudo-relevant documents and extracted from them the first 40 most informative terms which were added to the original query. The most informative terms are selected by using the information-theoretic Kullback-Leibler divergence function:

$$KL = f \cdot \log_2 \frac{f}{p} \tag{4}$$

where:

f is the term frequency in the set T,
p is the prior, e.g. the term frequency in the collection.

Once the computation of the Kullback-Leibler values KL are obtained and the new terms selected, we combine the initial term frequency of the term within the query (possibly equal to 0) with the score KL as follows:

$$tfq_{exp} = \frac{tfq}{Max_{tfq}} + \beta \cdot \frac{KL}{Max_{KL}} \tag{5}$$

where:

Max_{tfq} is the maximum number of occurrences of a term in the query
Max_{KL} is the highest KL value in the set T
β was set to 0.5

In the second retrieval we used the term weighting function:

$$w = tfq_{exp} \cdot (1 - Prob_1) \cdot (-log_2 Prob_2) \tag{6}$$

where $Prob_1$ and $Prob_2$ where defined as the first retrieval, that is according Formulas 2 and 3.

4 Augmenting PROSIT with word proximity

PROSIT, like most information retrieval systems, is based on index units consisting of single keywords, or unigrams, and it ignores word proximity. We attempted to improve its performance by using two-word index units (bigrams).

Bigrams are good for disambiguating terms and for handling topic drift, i.e., when the results of queries on specific aspects of wide topics contain documents that are relevant to the general topic but not to the requested aspect of it. This phenomenon can also be seen as some query terms matching out of context of their relationships to other terms [4]. For instance, using unigrams most returned documents for the CLEF query "Kaurismaki films" were about other famous films, whereas the use of bigrams considerably improved the precision of search.

On the other hand, some bigrams that are generated automatically may, in turn, over-emphasize concepts that are common to both relevant and nonrelevant documents [9]. So far, the results about the effectiveness of bigrams versus unigrams have not been conclusive.

We used a simple technique known as lexical affinities. Lexical affinities are identified by finding pairs of words that occur close to each other in a window of some predefined small size. For the CLEF experiments, we used the query title and chose a distance of 5 words. All the bigrams generated this way are seen as new index units and are used to increase the relevance of those documents that have the same pair of words occurring within the specified window. The score assigned to bigrams is computed using the same weighting function used for unigrams.

¿From an implementation point of view, in order to efficiently compute the bigram scores it is necessary to encode the information about the position of each term in each document into the inverted file. During query evaluation, for each bigram extracted from the query, the posting lists associated with the bigram words in the inverted file are merged and a new pseudo posting list is created that contains all documents that contain the bigram along with the relevant occurrence information.

The lexical affinity technique was reported to produce very good results on the web TREC collection, even better than those obtained using unigrams [5]. However, we were not able to obtain such good results on the CLEF collection, at least using queries with title and descriptions. In fact, we found that the bigram performance was considerably worse than the unigram performance; even when combining the scores, the performance remained lower than that obtainable by using just unigram scores.

Based on these observations, we decided to use bigrams in addition to, not in place of, single words. Second, instead of running two separate ranking systems, one for unigrams and the other for bigrams, and then combining their scores, we tried to incorporate the bigram component directly into PROSIT's main algorithm.

The bigram scores were thus combined with the unigram score to produce the first-pass ranking of PROSIT. In this way one can hope to increase the quality of the documents on which the subsequent query expansion step is based. This

may happen because more top relevant documents are retrieved or because the nonrelevant documents which contribute to query expansion are more similar to the query.

After the first ranking was computed using unigram and bigram scores, the top documents were used to generate the expanded query and PROSIT computed the second ranking as if it were just using unigrams. We chose to not expand the original query with two-word units due to the dimensionality problem, and we did not use the bigram method during the second-pass ranking of PROSIT because the order of words in the expanded query is not relevant.

We submitted one run to CLEF 2002, labeLLed as "fub02l", which was produced using PROSIT augmented with the bigrams procedure just described.

It should also be noted that we experimented with other types of multi-word index units, by using windows of different size and by selecting a larger number of words. However, we found that using just two words with a window of size 5 was the optimal choice.

5 Reranking PROSIT results using coordination level-based retrieval

Consistent with earlier effectiveness results, most information retrieval systems are based on best-matching algorithms between query and documents.

However, the use of very short queries on the part of most of the users and the prevailing interest in precision rather than recall have fostered new research on exact matching retrieval, seen as an alternative or as a complementary technique to traditional best-matching retrieval. In particular, it has been shown that taking into account the number of query words matched by the documents to rerank retrieval results may improve performance in certain situations (e.g.,[7], [3]).

For the CLEF experiments, we focused on the query title. The goal was to prefer documents that matched all of the query keywords above documents that matched all but one of the keywords, and so on.

To implement this strategy, we modified the standard best-matching similarity score between query and documents, computed as explained in Section 2, by adding a much larger addendum to it which was proportional to the number of distinct terms shared by the document and the query title. In this way, the documents were partially ordered according to their coordination level-based retrieval with the query title, with ties being broken using their best-matching similarity score to the query.

However, the results were somewhat disappointing. We obtained a much better retrieval effectiveness by simply preferring the documents that contained all the words of the query title, without paying attention to lower levels of coordination matching. This was our choice (run fub02b).

Finally, we submitted a fourth run by using the fully enhanced version of PROSIT, i.e., bigrams + coordination level-based retrieval (run fub02lb)

6 Results

We tested PROSIT and its three variants (i.e., PROSIT with bigrams, PROSIT with coordination level-based retrieval, and PROSIT with both bigrams and coordination level-based retrieval) on the CLEF 2001 and CLEF 2002 Italian monolingual tasks. Table 1 shows the retrieval performance of the four systems on the two test collections using the average precision as evaluation measure.

CLEF	PROSIT	PROSIT+bigrams	PROSIT+CLM	PROSIT+bigrams+CLM
CLEF 2001	0.5116	0.5208	0.5127	0.5223
CLEF 2002	0.5019	0.5088	0.4872	0.4947

Table 1. Retrieval performance of PROSIT and its variants.

The results of Table 1 show that the performance of standard PROSIT was excellent on both test collections, with the value obtained for CLEF 2001 (0.5116) being much higher than the result of the best system at CLEF 2001 (0.4865). This result is a confirmation of the high effectiveness of the probabilistic ranking model implemented in PROSIT, which is exclusively based on simple document and query statistics.

Table 1 also shows that, in general, the variations in performance when passing from basic PROSIT to enhanced PROSIT were small. More in particular, the use of bigrams improved performance across both test collections, whereas the use of coordination level-based retrieval was slightly beneficial for CLEF 2001 and detrimental for CLEF 2002.

Combining both enhancements improved the retrieval performance over using CLM alone on both test collections but it was still worse than baseline performance on the CLEF 2002 collection and worse than using bigrams alone on CLEF 2002.

Overall, the results about the enhanced versions of PROSIT are inconclusive. More work is needed to collect further evidence about their effectiveness, e.g., by using a more representative sample of performance measures or by considering other query scenarios.

Besides more robust evaluation of retrieval performance, it would be useful a better understanding of why the use of bigrams into PROSIT's main algorithm yielded positive results in the experiments reported in this paper. This might be done, for instance, by analysing the variations on quality of the top ranked documents used for query expansion or by performing a query by query analysis of concept drift in the final retrieved documents.

7 Conclusions

We have experimented with the PROSIT system on the Italian monolingual task and have explored the use of bigrams and coordination level-based retrieval within PROSIT's main algorithm. From our experimental evaluation, the following main conclusions can be drawn.

- The novel probabilistic model implemented in PROSIT achieved high retrieval effectiveness on both the CLEF 2001 and CLEF 2002 Italian monolingual tasks. These results are even more remarkable considering that the system employs very simple indexing techniques and does not rely on any specialised or ad hoc natural language processing techniques.
- Using bigrams in the place of unigrams hurt performance; the combination of bigram scores and unigram scores performed better but it was still inferior to the results obtained by using unigrams alone. However, using the bigram scores in the first-pass ranking, just to rank the documents used for query expansion, resulted in a performance improvement. These results held across both test collections.
- Using coordination level-based retrieval to rerank the retrieval results did not, in general, improve performance Favouring the documents that contained all the keywords in the query title worked better on one test collection and worse on the other collection, whereas ordering the documents according to their level of coordination matching hurt performance on both test collections.

We regret that due to tight schedule we were not able to test PROSIT on the other CLEF monolingual tasks. However, as the application of PROSIT to the Italian task did not require any special work, we are confident that with a small effort we could obtain similar results for the other languages. This is left for future work.

References

1. 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.
2. Gianni Amati and Cornelis Joost van Rijsbergen. Probabilistic models of information retrieval based on measuring divergence from randomness. *ACM Transactions on Information Systems*, (to appear), 2002.
3. E. Berenci, C. Carpineto, V. Giannini, S. Mizzaro. Effectiveness of keyword-based display and selection of retrieval results for interactive searches. *International Journal On Digital Libraries*, 3(3):249-260, 2000.
4. D. Bodoff, A. Kambil. Partial coordination. I. The best of pre-coordination and post-coordination. *JASIS*, 49(14):1254-1269,1998.

5. D. Carmel, E. Amitay, M. Herscovici, Y. Maarek, Y. Petruschka, A. Soffer, Juru at TREC-10 - Experiments with index pruning. In E.M. Voorhees and D.K. Harman, editors, *In Proceedings of the 10th Text Retrieval Conference TREC 2001*, pages 228-236, Gaithersburg, MD, 2002. NIST Special Pubblication 500-250.
6. 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.
7. C. Clarke, G. Cormack, E. Tudhope, (1997). Relevance ranking for one to three term queries. *Proceedings of RIAO'97*, 388-400, 1997.
8. J Savoy. Reports on clef-2001 experiments. In *Working notes of CLEF*, Darmstadt, 2001.
9. C.M Tan, Y.F. Wang, C.D Lee. The use of bigrams to enhance text categorization. *IP&M*, 38(4):529-546,2002.