The University of Lisbon at CLEF 2006 Ad-Hoc Task

Nuno Cardoso, Mário J. Silva and Bruno Martins
Faculty of Sciences, University of Lisbon
{ncardoso,mjs,bmartins}@xldb.di.fc.ul.pt

Abstract

This paper reports the participation of the XLDB Group from the University of Lisbon in the CLEF 2006 ad-hoc monolingual and bilingual subtasks for Portuguese. We present our IR system and detail both the query expansion strategy and the weighting scheme. In the end, we describe our runs and analyse our performance.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms

Design, Measurement, Experimentation

Keywords

Information Retrieval, Evaluation, CLEF, ad-hoc

1 Introduction

This paper describes the third participation of the XLDB Group from the University of Lisbon in the CLEF ad-hoc Portuguese monolingual task and English to Portuguese bilingual task. Our main goal is to obtain a stable platform to test GIR approaches for the GeoCLEF task [1].

In 2004 we participated with an IR system made from components of tumba!, our web search engine [2]. We learnt that searching and indexing large web collections is different than querying CLEF ad-hoc newswire collections [3]. Braschler and Peters overviewed the best IR systems of the CLEF 2002 campaign and concluded that they relied on robust stemming, good term weighting scheme and a query expansion approach [4]. Since tumba! does not have a stemming module nor a query expansion module, and its weighting scheme is built for web documents and based on PageRank [5] and in HTML markup elements, we needed to develop new modules to handle properly the adhoc task.

In 2005 we developed QuerCol, a query generator module with query expansion, and we implemented a tf$\times$idf term weighting scheme with a result set merging module for our IR system [6]. The results improved, but were still far from our expectations. This year we improved QuerCol with a blind relevance feedback algorithm, and implemented a term weighting scheme based on BM25 [7].

The rest of this paper is organised as follows: in Section 2 we describe our IR system and detail the improvements made to QuerCol and SIDRA, our indexing, retrieval and ranking module. In Section 3 we present our evaluation goals and submitted runs. Section 4 shows the results obtained, and Section 5 summarises our conclusions and refers future work.

2 IR system architecture

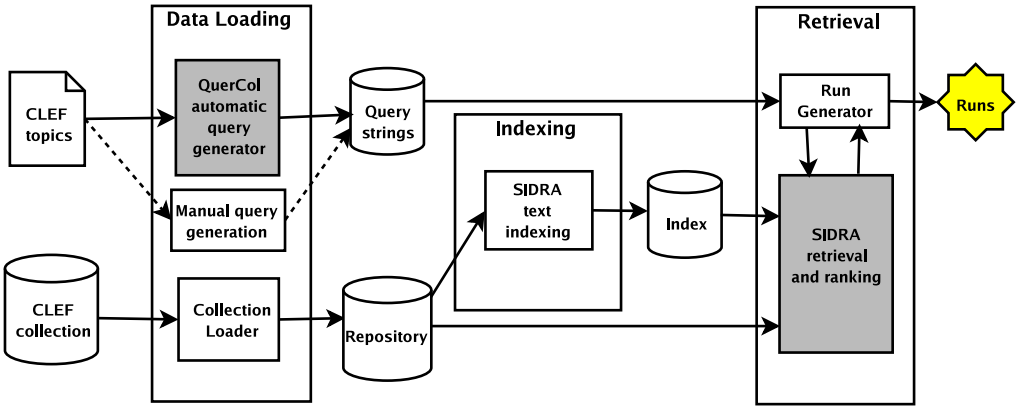

Figure 1: Our IR system architecture.

Figure 1 presents our IR system architecture. In the data loading step, the CLEF collection is loaded in a repository, so that SIDRA can index the collection and generate term indexes in the indexing step. For our automatic runs, QuerCol loads the CLEF topics and generates query strings. In the retrieval step, the queries are submitted to SIDRA through a run generator, producing runs in CLEF format. In the rest of this Section we detail the modules shadowed in grey, QuerCol and SIDRA.

2.1 QuerCol query generator

QuerCol (Query Collator) was first developed in 2005 for our ad-hoc participation. Quercol processed the CLEF topics for our IR system, combining the topic terms with their morphological derivatives to produce boolean query strings in Disjunctive Normal Format, that is, with disjunctive groups of conjunctive terms [6]. The results were promising but insufficient, so we reformulated QuerCol with a query expansion step using blind relevance feedback [8, 9]. Together with a query construction step, QuerCol can parse CLEF topics and generate query strings without human intervention. QuerCol is detailed on Figure 2, and can be resumed in three stages:

Stage 1: Initial run generation

.

For each topic, the non-stopword terms from the title are extracted and combined into a boolean query with AND operators. The initial query is submitted to SIDRA, generating an initial run. Note that, in our automatic runs, we did not used the description and the narrative fields.

Stage 2: Term ranking

We used the $w_t(p_t\text{-}q_t)$ algorithm to weight the terms for our query expansion algorithm, as represented in the following equation [10]:

$$a_t = w_t\,(p_t - q_t) = log\frac{(r+0.5)\,(N-n-R+r+0.5)}{(n-r+0.5)\,(R-r+0.5)} \times \left(\frac{r}{R} - \frac{n-r}{N-R}\right)$$

where N is the total number of documents, n is the number of documents containing the term t, R is the number of relevant documents, r is the number of relevant documents containing the term t. In the end, p_t represents the probability of the term t to occur in a relevant document, while q_t represents the probability of the term t to occur in a non-relevant document.

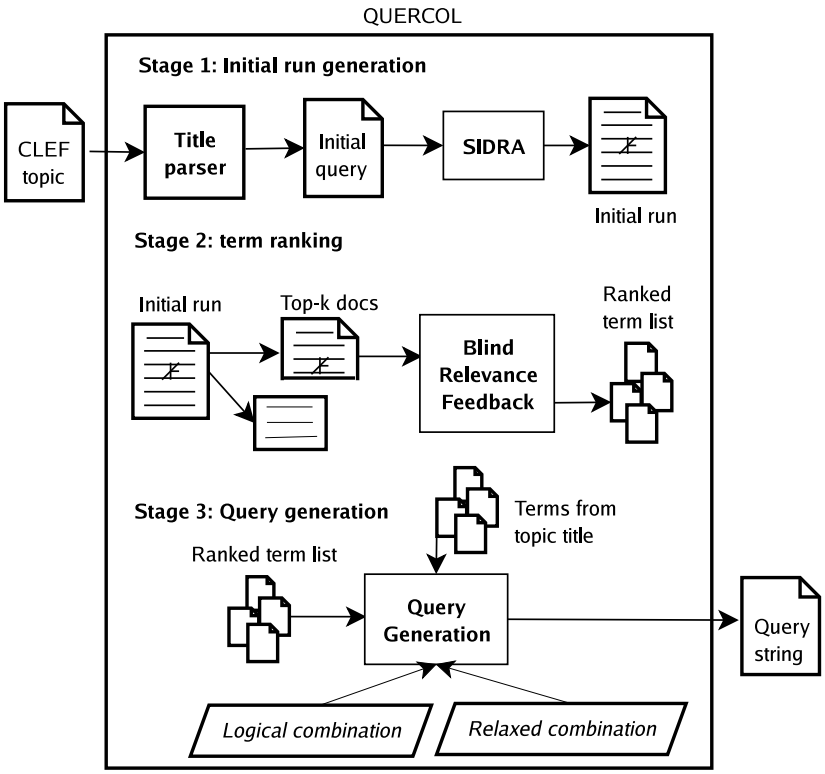

Figure 2: Details of the QuerCol module.

QuerCol assumes that all documents above a certain threshold parameter, the *top-k documents*, are relevant for a given topic, and that the remaining documents are non-relevant. The top-k documents are then tokenised and their terms are weighted, generating a *ranked term list* that best represent the top-k documents, and thus the information need of the topic.

Stage 3: Query generation

Finally, QuerCol generates a boolean query string in the Disjunctive Normal Format, using terms from the ranked term list and from the topic title. The terms are combined using one of the two following query construction approaches, the *Logical combination* and the *Relaxed combination*.

The *Logical combination* assumes that all non-stopwords from the topic title represents different *concepts*, and that these concepts must be mentioned in the retrieved documents (see Figure 3). The queries generated by the Logical combination reproduces the boolean constraints described by Mitra et al [11], forcing each query to contain at least one term from each concept.

As each concept may be represented by many terms, the Logical combination approach searches the ranked term list to find related terms for each concept. When found, the terms are moved into the corresponding concept's bag of terms, the *concept bags*. QuerCol relates the term to a concept if they share the same stem, as given by Porter's stemming algorithm for Portuguese [12]. After filling all concept bags, the remaining top-k terms from the ranked term list are moved into a new bag, called *expanded terms bag*. This bag contains terms that are strongly related to the topic, but do not share the same stem from any of the concepts.

The query generation step produces all possible combinations (${}^{n}C_{m}$) of the $m \times n$ matrix of m bags $\times$ n terms in each bag, resulting in $m \times n$ *partial queries* containing one term from each concept bag. The terms of the partial queries are joined by the AND operator, with the exception of the terms from the expanded terms bag, that are joined by the OR operator. We made this exception for the expanded terms to avoid query drift. In the Disjunctive Normal Format, the final query string is the following:

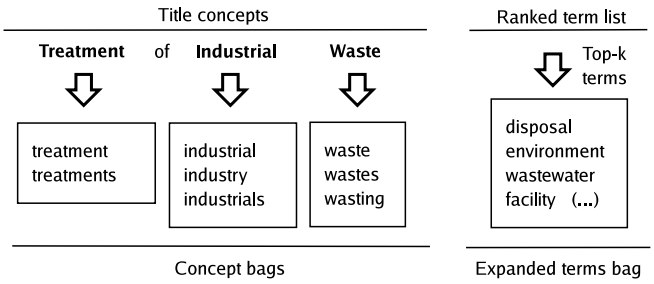

Figure 3: The Logical combination approach for query generation.

OR{ OR{*partial queries* $_{\text{AND\{concept bags\}}}$}, OR{*partial queries*$_{\text{AND\{concept bags + expanded terms bag\}}}$ } }

Nonetheless, Mitra et al pointed that there are relevant documents that may not mention all concepts from the topic title [11]. As the Logical combination forces all concepts to appear in the query strings, we may not retrieve several relevant documents. Indeed, QuerCol generated query strings in a Logical combination approach for our 2005 ad-hoc participation, and we observed low recall values in our results [6].

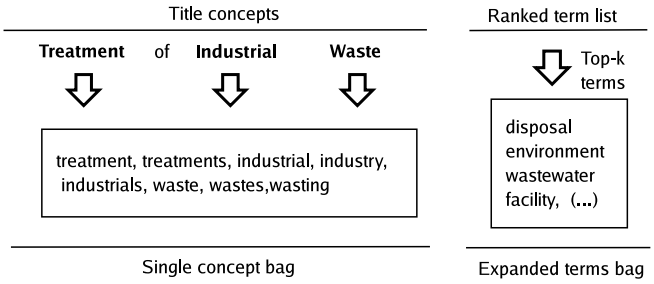

Figure 4: The Relaxed combination approach for query generation.

To tackle the limitations of the Logical combination, we implemented a modified version of the Logical combination, called Relaxed combination. The Relaxed Combination differs from the the Logical combination by using a single bag to collect the related terms (the *single concept bag*), instead of a group of concept bags (see Figure 4). This modification relaxes the boolean constraints from the Logical combination.

The ${}^{n}C_{m}$ combinations of the two bags (the single concept bag and the expanded terms bag) use the AND operator, producing $m \times n$ partial queries. In the Disjunctive Normal Format, the final query string is the following:

OR{ *partial queries*$_{\text{AND\{single concept bag + expanded terms bag\}}}$}

2.2 Weighting and Ranking

The SIDRA retrieval and ranking module implemented the BM25 weighting scheme, where the weight for each term is computed as follows:

$$BM25(t_i) = \frac{(k_1+1) \times term\_freq(t_i)}{k_1 \times ((1-k_2) + k_2 \times \frac{doc\_len}{avg\_doc\_len}) + d} \log(\frac{N - doc\_freq(t_i) + 0.5}{doc\_freq(t_i) + 0.5})$$

The parameters were set to the standard values of k_1=2.0 and k_2=0.75.

Robertson et al. have recently proposed an extension of the BM25 scheme for structured documents, suggesting that document elements such as the title could be repeated in a corresponding unstructured document, so that the title terms can be weighted more important [13].

For CLEF, we assumed that the first three sentences of each document should be weighted as more important, as the first sentences of news articles often contains a summary of the content. Robertson's extension was applied in run PT4, giving a weight of 3 to the first sentence, and a weight of 2 to the following two sentences.

SIDRA's ranking module was also improved to support disjunctive queries more efficiently, so we discarded the result merging set module that we used in CLEF 2005 [6].

3 Runs

We submitted 4 runs for the Portuguese ad-hoc monolingual subtask and 4 runs for the English to Portuguese ad-hoc bilingual subtask. The Portuguese monolingual runs evaluated both QuerCol query construction strategies and the BM25 term weigh extension, while the English runs evaluated different values for the top ranked document threshold (top-k documents), and for the size of the expanded terms bag (top-k terms). Table 1 resumes the configuration of the submitted runs.

Table 1: Description of the submitted runs for the Portuguese monolingual subtask (PT) and the English to Portuguese bilingual subtask(EN).

Label	Type	Query construction	top-k terms	top-k docs	BM25 extension
PT1	Manual	Manual	-	20	no
PT2	Automatic	Logical	8	20	no
PT3	Automatic	Relaxed	32	20	no
PT4	Automatic	Relaxed	32	20	yes
EN1	Automatic	Relaxed	16	10	no
EN2	Automatic	Relaxed	32	10	no
EN3	Automatic	Relaxed	16	20	no
EN4	Automatic	Relaxed	32	20	no

The PT1 run was created manually from topic terms, their synonyms and morphological expansions. For our automatic runs, we used the CLEF 2005 topics and qrels to find the best top-k term values for a fixed value of 20 top-k documents. The Logical combination obtained a maximum MAP value of 0.2099 for 8 top-k terms. The Relaxed combination did not performed well for low top-k term values, but for higher values it outperformed the Logical combination with a maximum MAP value of 0.2520 for 32 top-k terms.

For the English to Portuguese bilingual subtask, we translated the topics with Babelfish (`http://babelfish.altavista.com`), and used different values for top-k terms and top-k documents, to evaluate if they significantly affect the results.

4 Results

Table 2 presents our results. For the Portuguese monolingual subtask, we observe that our best result was obtained by the manual run, but the automatic runs achieved a performance comparable to the manual run. The Relaxed combination produced better results than the Logical combination, generating our best automatic runs. The BM25 extension implemented in the PT4 run did not produce significant improvements.

For the English to Portuguese bilingual task, we observe that the different top-k terms and top-k document values do not affect significantly the performance of our IR system. The PT3 and EN4 runs were generated with the same configuration, to compare our performance in both subtasks. The monolingual run obtained the best result, with a difference of 32% in the MAP value to the corresponding bilingual run.

Table 2: Overall results for all submitted runs.

Measure	PT1	PT2	PT3	PT4	EN1	EN2	EN3	EN4
num_q	50	50	50	50	50	50	50	50
num_ret	13180	7178	48991	49000	41952	42401	42790	43409
num_rel	2677	2677	2677	2677	2677	2677	2677	2677
num_rel_ret	1834	1317	2247	2255	1236	1254	1275	1303
map	0,3644	0,2939	0,3464	0,3471	0,2318	0,2371	0,2383	0,2353
gm_ap	0,1848	0,0758	0,1969	0,1952	0,0245	0,0300	0,0377	0,0364
R-prec	0,4163	0,3320	0,3489	0,3464	0,2402	0,2475	0,2509	0,2432
bpref	0,3963	0,3207	0,3864	0,3878	0,2357	0,2439	0,2434	0,2362
recip_rank	0,7367	0,7406	0,6383	0,6701	0,4739	0,4782	0,5112	0,4817
ircl_prn.0.00	0,78	0,77	0,69	0,71	0,52	0,52	0,55	0,53
ircl_prn.0.10	0,65	0,63	0,57	0,57	0,4	0,4	0,4	0,39
ircl_prn.0.20	0,59	0,5	0,51	0,5	0,33	0,34	0,34	0,34
ircl_prn.0.30	0,51	0,43	0,45	0,44	0,28	0,31	0,28	0,29
ircl_prn.0.40	0,46	0,35	0,4	0,39	0,26	0,27	0,26	0,27
ircl_prn.0.50	0,39	0,28	0,35	0,35	0,23	0,23	0,23	0,24
ircl_prn.0.60	0,34	0,19	0,3	0,3	0,21	0,2	0,2	0,21
ircl_prn.0.70	0,2	0,13	0,27	0,27	0,18	0,18	0,18	0,19
ircl_prn.0.80	0,14	0,08	0,2	0,21	0,15	0,14	0,16	0,14
ircl_prn.0.90	0,08	0,05	0,14	0,13	0,09	0,09	0,1	0,08
ircl_prn.1.00	0,02	0,01	0,08	0,08	0,07	0,07	0,08	0,06
P5	0,58	0,54	0,5	0,5	0,33	0,35	0,34	0,36
P10	0,54	0,5	0,48	0,48	0,31	0,32	0,33	0,32
P15	0,5	0,46	0,45	0,43	0,29	0,3	0,3	0,29
P20	0,47	0,43	0,42	0,41	0,27	0,28	0,28	0,27
P30	0,42	0,38	0,37	0,37	0,23	0,25	0,25	0,24
P100	0,25	0,2	0,23	0,22	0,13	0,13	0,14	0,14
P200	0,15	0,11	0,15	0,15	0,09	0,09	0,09	0,09
P500	0,07	0,05	0,08	0,08	0,05	0,05	0,05	0,05
P1000	0,04	0,03	0,04	0,05	0,02	0,03	0,03	0,03

5 Conclusion

For our participation on the CLEF 2006 adhoc task, we implemented well-known algorithms in our IR system to obtain good results on the ad-hoc task, allowing us to stay focused on GIR approaches for the GeoCLEF task.

Our results show that we improved our monolingual IR performance in precision and in recall. The best run was generated from a query built with a Relaxed combination, with an overall recall value of 84.2%. We can not tell at this time what are the contributions of each module to the achieved improvements of the results.

The English to Portuguese bilingual results show that the topic translation was poor, resulting in a decrease of 0.111 in the MAP values for runs PT3 and EN4. The difference between there two runs show that we need to implement a specific translation module to improve our bilingual results.

We also observe that the top-k terms and top-k document parameters did not affect significantly the performance of the IR system.

Next year, our efforts should focus on improving the query expansion and query construction algorithms. QuerCol can profit from the usage of the description and narrative fields, producing better query strings. Also, we can follow Mitra et al. suggestion and rerank the documents before the relevance feedback, to ensure that the relevant documents are included in the top-k document set [11]. An additional relevance repository that can be exploited is the tumba! logs, to infer relationships between query terms and document contents, in a similar way to what Fitzpatrick et al suggested. [14]

6 Acknowledgements

We would like to thank Daniel Gomes, who built the tumba! repository used on our participation, and to all developers of the tumba! search engine. Thanks also to Alberto Simões for providing the topic translations. Our participation was partly financed by the Portuguese Fundação para a Ciência e Tecnologia through grants POSI / PLP / 43931 / 2001 (Linguateca) and POSI / SRI / 40193 / 2001 (GREASE).

References

[1] Martins, B., Cardoso, N., Chaves, M., Andrade, L., Silva, M.J.: The University of Lisbon at GeoCLEF 2006. In Peters, C., ed.: Working Notes for the CLEF 2006 Workshop, Alicante, Spain (2006)

[2] Silva, M.J.: The Case for a Portuguese Web Search Engine. In: Proceedings of ICWI-03, the 2003 IADIS International Conference on WWW Internet, (Algarve, Portugal) 411–418

[3] Cardoso, N., Silva, M.J., Costa, M.: The XLDB Group at CLEF 2004. In Peters, C., Clough, P., Gonzalo, J., Jones, G., M.Kluck, Magnini, B., eds.: Proceedings of the 5th Workshop of the Cross-Language Evaluation Forum, CLEF'2004. Volume 3491 of Lecture Notes in Computer Science., Bath, UK, Springer (2005) 245–252

[4] Braschler, M., Peters, C.: Cross-language evaluation forum: Objectives, results, achievements. Information Retrieval **7** (2004) 7–31

[5] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library (1999)

[6] Cardoso, N., Andrade, L., Simões, A., Silva, M.J.: The XLDB Group participation at CLEF 2005 ad hoc task. In Peters, C., Clough, P., Gonzalo, J., Jones, G., M.Kluck, Magnini, B., eds.: Proceedings of the 6th Workshop of the Cross-Language Evaluation Forum, CLEF'2005. Volume 4022 of Lecture Notes in Computer Science., Springer-Verlag (2006) 54–60

[7] Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M.: Okapi at TREC-3. In IST Special Publication 500-225 Harman, D., ed.: Overview of the Third Text REtrieval Conference (TREC 3), Gaithersburg, MD, USA, Department of Commerce, National Institute of Standards and Technology (1995) 109 – 126

[8] Rocchio Jr., J.J.: Relevance feedback in information retrieval. In Salton, G., ed.: The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall, Englewood Cliffs, NJ, USA (1971) 313–323

[9] Efthimiadis, E.N.: Query expansion. In Williams, M.E., ed.: Annual Review of Information Systems and Technology (ARIST). Volume 31. (1996) 121–187

[10] Efthimiadis, E.N.: A user-centered evaluation of ranking algorithms for interactive query expansion. In: Proceedings of ACM SIGIR '93. (1993) 146–159

[11] Mitra, M., Singhal, A., Buckley, C.: Improving automatic query expansion. In: Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, ACM Press (1998) 206–214

[12] Porter, M.: An algorithm for suffix stripping. Program **14** (1980) 130–137

[13] Robertson, S., Zaragoza, H., Taylor, M.: Simple BM25 extension to multiple weighted fields. In: CIKM '04: Proceedings of the thirteenth ACM international Conference on Information and Knowledge Management, New York, NY, USA, ACM Press (2004) 42–49

[14] Fitzpatrick, L., Dent, M.: Automatic feedback using past queries: Social searching? In: Proceedings of the 20th International Conference on Research and Development in Information Retrieval (SIGIR-97), Philadelphia, USA (1997) 306–313