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, detail the query expansion strategy and the weighting scheme, describe the submitted runs and discuss the obtained results.

1 Introduction

This paper describes the third participation of the XLDB Group from the University of Lisbon in the CLEF ad-hoc task. Our main goal was to obtain a stable platform to test GIR approaches for 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, a good term weighting scheme and a query expansion approach [4]. Tumba! does not have a stemming module and does not perform query expansion. Its weighting scheme, built for web documents, is based on PageRank [5] and in HTML markup elements. As a result, we needed to develop new modules to properly handle the ad-hoc task.

In 2005 we developed QuerCol, a query generator module with query expansion, and 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 performance goal. This year we improved QuerCol with a blind relevance feedback algorithm, and implemented a term weighting scheme based on BM25 [7].

2 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, the indexing system of tumba!, can index the collection and generate term indexes. 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.

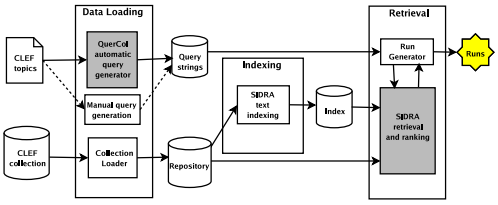

Fig. 1. The IR system architecture.

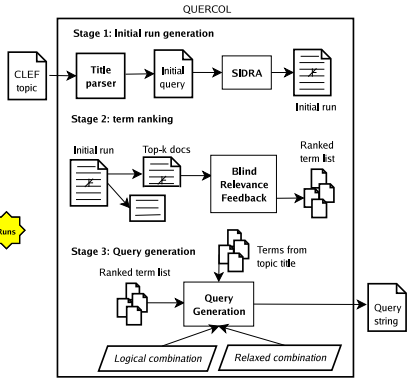

Fig. 2. Details of the QuerCol module.

2.1 QuerCol query generator

This year, we improved 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 operates in three stages (see Figure 2):

Stage 1: Initial run generation For each topic, the non-stopword terms from its title are extracted and combined as Boolean AND expressions, yielding the queries submitted to SIDRA that generated the initial runs. Note that, in our automatic runs, we did not use the description or 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 [10]. QuerCol assumes that only the documents above a certain threshold parameter, the *top-k documents*, are relevant for a given topic. The top-k documents are then tokenised and their terms are weighted, generating a *ranked term list* that best represents the top-k documents.

Stage 3: Query generation QuerCol combines terms from the ranked term list and from the topic title, to generate a Boolean query string in the Disjunctive Normal Format (DNF). This combination may use the AND expression (operation $\times_{AND}$) or the OR expression (operation $\times_{OR}$). We defined two query construction approaches, the *Logical Combination* (LC) and the *Relaxed Combination* (RC).

The *Logical Combination* assumes that all non-stopwords from the topic title are different *concepts* that must be mentioned in the retrieved documents (see Figure 3). The generated query string forces each query to contain at least one term from each concept, reproducing the Boolean constraints described in [11].

As each concept may be represented by many terms, the LC approach searches the ranked term list to find related terms for each concept. When found, the terms are moved

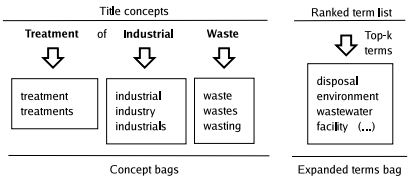

Fig. 3. Logical Combination (LC) approach.

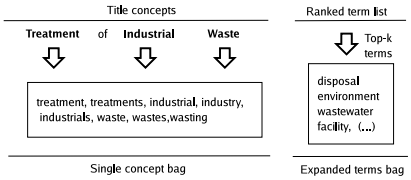

Fig. 4. Relaxed Combination (RC) approach.

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 next stage generates all possible combinations of the $m \times n$ matrix of m bags $\times$ n terms in each bag using the the $\times_{AND}$ operation, producing $m \times n$ *partial queries* containing one term from each content bag and from the expanded terms bags. The partial queries are then combined with the $\times_{OR}$ operation, resulting in a query string ($\times_{OR}[partial\ queries_{\text{concept + expanded terms bags}}]$). We found that these query strings were vulnerable to query drift, so we generated an additional query string using only the concept bags in a similar way ($\times_{OR}[partial\ queries_{\text{concept bags}}]$), and then combined both query strings using an $\times_{OR}$ operation. In DNF, the final query string generated by the LC approach is the following:

$$\times_{OR}\left[\times_{OR}[partial\ queries_{\text{concept bags}}], \times_{OR}[partial\ queries_{\text{concept + expanded terms bags}}]\right]$$

Nonetheless, there are relevant documents that may not mention all concepts from the topic title [11]. As the LC forces all concepts to appear in the query strings, we may not retrieve some relevant documents. Indeed, last year QuerCol generated query strings following an LC approach, and we observed low recall values in our results [6].

To tackle the limitations of the LC, we implemented a modified version, called the *Relaxed Combination*. The RC differs from the the LC 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 LC. The RC approach generates partial queries in a similar way as the LC approach, combining terms from the two bags (the single concept bag and the expanded terms bag) using the $\times_{AND}$ operation. The partial queries are then combined using the $\times_{OR}$ operation, to generate the final query string. In DNF, the RC approach is the following:

$$\times_{OR}\left[partial\ queries_{\text{single concept + expanded terms bags}}\right]$$

2.2 Weighting and Ranking

The SIDRA retrieval and ranking module implemented the BM25 weighting scheme. The parameters were set to the standard values of k_1=2.0 and k_2=0.75. Robertson et al. 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 to generate 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 abandoned the result sets merging module that we developed for CLEF 2005 [6].

3 Results

Table 1. Runs submitted for the Portuguese (PT) and English (EN) monolingual.

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

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

Run PT1 was manually created 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 LC runs obtained a maximum MAP value of 0.2099 for 8 top-k terms. The RC runs did not perform well for low top-k term values, but for higher values they outperformed the LC runs 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 tested with half of the top-k terms and top-k documents, to evaluate if they significantly affect the 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 RC produced better results than the

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

LC, 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.

4 Conclusion

This year, 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 both precision and 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 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 the two runs shows that we need to adopt another strategy to improve our bilingual results. We also observe that the top-k terms and top-k document values 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 the suggestion of Mitra et al. and rerank the documents before the relevance feedback, to ensure that the relevant documents are included in the top-k document set [11].

Acknowledgements We would like to thank Daniel Gomes who built the tumba! repository, Leonardo Andrade for developing SIDRA, Alberto Simões for topic translations, and to all developers of tumba!. Our participation was partially supported by grants POSI/PLP/43931/2001 (Linguateca) and POSI/SRI/40193/2001 (GREASE) from FCT (Portugal), co-financed by POSI.

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. **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