The XLDB Group at CLEF 2004

Nuno Cardoso, Mário J. Silva and Miguel Costa
Grupo XLDB - Departamento de Informática
Faculdade de Ciências da Universidade de Lisboa
{ncardoso, mjs, mcosta} at xldb.di.fc.ul.pt

12th August 2004

Abstract

This paper describes the participation of the XLDB Group in the Monolingual IR task for the Portuguese language. We present tumba!, a Portuguese search engine, and we describe its architecture and asumptions. We discuss the way we used tumba! in CLEF, detailing the submitted runs and our experiments with ranking algorithms. In the end, we evaluate our participation and comment on future work.

1 Introduction

In 2004, for the first time, CLEF included Portuguese document collections for Monolingual & Bilingual Information Retrieval and Question Answering tasks. This collection [14] was based on news of several categories taken from Publico [13], a Portuguese newspaper, and compiled by Linguateca [3]. This year, the XLDB Group made its debut participation in CLEF.

This paper is organized as follows: in Section 2, we introduce the XLDB Group. In section 3, we describe tumba!, our IR system, and the modifications we made to it to handle the CLEF 2004 data set. Section 4 describes the official runs with the implemented algorithms for CLEF 2004, and Section 5 presents our results. 6 summarizes a conclusion of our participation.

2 The XLDB Group

The XLDB Group is a research unit of LaSIGE (Large Scale Information Systems Laboratory) at FCUL - Faculdade de Ciências da Universidade de Lisboa. We research data management systems for data analysis, information integration and user access to large quantities of complex data from heterogeneous platforms. Current research lines span Web search, mobile data access, temporal web data management and bioinformatics.

The XLDB Group is involved in several projects and activities. One of our main projects is tumba! [8, 15], a Portuguese Web search engine. Tumba! is described in Section 3.

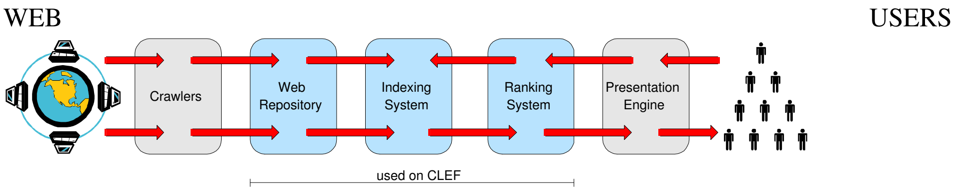

Figure 1: tumba's architecture

The XLDB Group hosts a node of Linguateca, a distributed language resource center for Portuguese, since January 2004 [7].

The participation of the XLDB Group in the Monolingual Task for Portuguese language, with the tumba! search engine, was motivated by several reasons:

1. Although we had previous experiences in evaluation contests, namely in the biotext task of the KDD Cup 02 [5] and in the BioCreative workshop [6], this was our first opportunity for evaluating tumba! jointly with other IR systems, with the advantage of the evaluation being conducted on a Portuguese collection.

2. Although we were aware that our system was out of its natural environment, the Web, we could take the opportunity to tune the indexing and ranking engines of tumba!, by submitting our results using different ranking configurations and then analyzing the results.

3. To gather ideas on how to improve our search engine results.

3 Tumba in the Monolingual Task

3.1 Overview of tumba!

Tumba! is a search engine specially crafted to archive and provide search services to a community Web formed by those interested in subjects related to Portugal and the Portuguese people [8]. Tumba! is being offered as a public service since November 2002.

Tumba is mainly written in Java and built in open-source software, such as the Linux operating system. It has an index of over 3.5 million Web documents and a daily traffic of up to 20.000 queries per day. Its response time is less than 0.5 seconds for 95% of the requests. It is also a platform for PhD and MSc research projects at our university.

Tumba! has a similar architecture to global search engines and adopts many of the algorithms used by them [1]. However, its configuration data is much richer in its domain of specialisation. Tumba! has a better knowledge of the location and organization of Portuguese Web sites (both in qualitative and quantitative terms) [15].

The data flows from the Web to the user through a pipeline of the following tumba! sub-systems (See Figure 1):

Crawlers: collect documents from the Web, given an initial URL list. They parse and extract URLs from each document, which will be used to collect new documents. These steps are performed recursively until a stop condition is met [10].

Web Repository: The Web data collected by the crawlers is stored in Versus, a repository of Web documents and associated meta-data [9].

Indexing system: the indexing system Sidra creates indexes over the documents in the Web Repository [4], so that when a query is received, Sidra uses the indexes built to find the documents that match that query.

Ranking system: computes, for each document *d* returned by the indexing system, a similarity value between *d* and the submitted query using a set of heuristics. Then, it sorts the documents by these similarities.

Presentation Engine: formats the result sets received from the ranking engine for the user's access platforms such as Web browsers, PDA devices or WAP phones.

3.2 Portuguese Monolingual Task

The previous CLEF tasks showed that the top performing groups for Monolingual IR tasks were systems which performed robust stemming, well-known weighting schemes (BM25, Lnu.ltn or Berkeley ranking) and blind feedback or query expansion [12]. Tumba's system doesn't have a stemmer and a blind feedback or query expansion system, and the term weighting scheme is tuned for Web searches. Still, we decided that tumba! should suffer no architectural change to be used in this evaluation. We wanted to evaluate tumba!'s performance with its current components, so that we could have a baseline for comparison on future CLEF tasks. Nonetheless, we felt that our participation in CLEF would provide us with valuable ideas to optimize our search engine results, and resources to evaluate our system performance.

One of the difficulties we encountered on the CLEF Monolingual task was related to the SGML-format used on collection of Portuguese documents. The documents have tags for associated metadata like author, category and date of publication. The contents are in plain text, with no additional tags. Tumba! was not conceived to work with document collections organized like this. Its ranking system was developed to profit from annotations extracted from the Web documents, such as:

- Information obtained from the Web graph, like links and anchor text, which are a valuable resource to find related pages that might interest the user;
- Documents' structural elements like titles and headings, which provide valuable information of the document subject.

We used the same alghorithms designed for the Web in CLEF, despite the different search context. The lack of this kind of "light semantic" annotation in the collection was a major handicap for the tumba! system, since the only semantic information we managed to extract from the documents was the news' titles. Our heuristic for extracting documents? titles consisted in finding paragraphs in the collection with a maximum of 15 terms and ending with no punctuation.

We disabled the query-independent ranking calculations and most of the emphasis ranking augmenters of the Indexing and Ranking system, since there wasn't such information on the collection.

Tumba's Crawlers and Presentation Engine weren't used for the CLEF Portuguese Monolingual IR task. We loaded the document collection directly into the Web Repository, bypassing the system's crawlers. The collection was then indexed by the Sidra Indexing system. Queries were sent directly to Sidra, bypassing the Presentation Engine, and the matching documents were then ranked according to some heuristics to compute document relevance.

4 Runs

The Monolingual IR task limited, for groups in their first participation, the number of submitted runs to 4.

4.1 Manual Run (XLDBTumba01)

Since this was the first time that CLEF used Portuguese collections in an evaluation campaign, this task didn't have previous relevant judgements and training collections. In order to have a prior evaluation of tumba!, we created our own baseline against which we could compare our runs to measure how much we were improving our system.

For each one of the 50 given topics, we created several different queries related to the topic and we used them to retrieve documents matching the query terms. Then, the returned results were manually examined by two doctoral students, with some IR systems usage experience but unfamiliar with the tumba! system, and classified the documents as relevant or irrelevant according to the topic criteria. This was a laborious work, which consumed most of the time for this task.

After that, we compiled a list of the relevant documents and submitted it to CLEF as our run XLDBTumba01, to measure the offset of our baseline compared to the CLEF solutions.

When the relevant judgements were released by CLEF, we observed that we had many errors in our manual review; from incorrect topic interpretation to bad query formulation. In the end, this was the run that had the worst performance. Yet, this run clearly showed to us how difficult it is to formulate queries that correctly match an information need.

4.2 Flat Ranking Run (XLDBTumba02)

For subsequent runs, we chose among the different queries used to create the XLDB-Tumba01 run to select which 50 queries would be used on the remaining runs. Note that we didn't use more than one query per topic, neither did any kind of query expansion.

This run was produced by submitting the 50 queries directly to the Sidra Indexing and Ranking system, configured to perform an exact matching (flat-ranking algorithm), returning only the documents that match all the query terms.

We see this run as our automatic baseline run, and we were anticipating that the other runs would improve precision and recall compared to this run. Yet, this run outperformed all the other runs.

4.3 Distances Run (XLDBTumba05)

This run was generated using the following ranking algorithm:

- *distMinTerms(d,q)* - uses the minimum distances between any pair of query terms *q* in documents *d*, *minDist*, to increase the ranking of documents whose query terms are closer on the document. For distances above 10, the function gives similarity 0 to the document. If all query terms are adjacent on a document, their *minDist* value equals 1.

$$distMinTerms(d,q) = \begin{cases} 1 & minDist = 1 \\ 1 - \frac{minDist-1}{9} & 1 < minDist < 10 \\ 0 & minDist \geq 10 \end{cases}$$

This function indeed improved the results accordingly to our own evaluation, as the queries with more than one term we used for the topic tend to be adjacent.

4.4 Distances + Titles Run (XLDBTumba04)

This run was generated by using two ranking algorithms in Sidra:

- *distMinTerms(d,q)*
- *termsInTitle(d,q)* - this is a similarity function between the terms in the title of each document *d*, denoted *T*, and the query terms in a query *q*, denoted *Q*.

$$termsInTitle(d,q) = \frac{|T \cap Q|}{max(|T|,|Q|)}$$

This run evaluated the importance of the title in the document ranking, and turned out as the one with the worst performance in our self-evaluation. This was probably caused by the heuristic used to extract titles from the documents, which was a very naive approach and may have mislead the ranking engine. The tumba! search engine gives great importance to title texts, as many people search named entities on search engines and these are usually clearly stated in the titles.

5 Results

For a prior evaluation of our automatic runs, we compared the results with manual run XLDBTumba01. We used precision@1, precision@3, precision@10, recall and F-Measure ($\beta = 1$) metrics in our self-evaluation. The results are summarized on Table 1.

Run	Description	Precision@			Recall	F-Measure
		1	3	10		
XLDBTumba02	flat ranking	53.2%	47.2%	40.6%	89.6%	44.4%
XLDBTumba05	Distances	46.8%	53.5%	44.9%	89.6%	44.4%
XLDBTumba04	Distances & Titles	48.9%	45.0%	41.1%	89.6%	44.4%

Table 1: Automatic Submitted Runs, compared to the Manual Run XLDBTumba01

The results obtained in CLEF are presented on Table 2 and Figure 2. The Average Precision (non-interpolated) for all relevant documents and the R-Precision (precision after R documents retrieved) are the measures presented by the trec_eval program. [2, 11]

Run	Manual Run	Flat ranking	Distances	Distances + titles
	(XLDBTumba01)	XLDBTumba02	XLDBTumba05	XLDBTumba04
Nr. Docs Retrieved	209	2350	2350	2350
Nr. Relevant Docs	678	678	678	678
Relevant Docs Retrieved	79	168	168	168
Overall Precision	37,8%	7,1%	7,1%	7,1%
Overall Recall	11,6%	24,8%	24,8%	24,8%
Average Precision	21,84%	28,10%	25,13%	27,75%
R-Precision	22,41%	26,28%	26,73%	27,26%

Table 2: XLDB official runs evaluated by CLEF

The XLDBTumba02, XLDBTumba05 and XLDBTumba04 runs have the same overall precision and recall values because we used the same queries which retrieved the same documents, differing only in the order on which the documents were submitted for each topic.

6 Conclusion

We used the Web search engine tumba! in the CLEF 2004 Monolingual task for the Portuguese language. Our main objective was to test, compare, and improve the quality of tumba's results, and gather ideas on how to do it. However, the enviroment that we work on, the Web, is different from the flat and small collection of document texts that we used on the CLEF task.

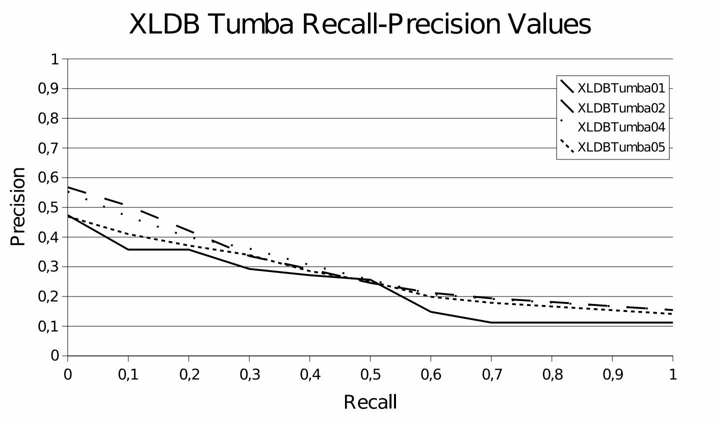

Figure 2: Recall-Precision Values for our runs, according to CLEF results.

As we didn't had a baseline of relevant judgements, we manually annotated relevant and non relevant documents for the 50 topics. We found that this task isn't easy, as it is time consuming and it requires experienced human annotators to review hundreds of documents, cross the results and eliminate erroneous judgements. The other submitted runs used combinations of two algorithms used on the tumba! ranking engine. We did our own evaluation with several metrics based on our own relevance judgements, and we submitted 4 runs for CLEF evaluation. We presented both evaluations in this paper.

Tumba! does not perform stemming or query expansion and relies heavily on detecting the presence of query terms in document titles and URLs. As these were not available for this evaluation, our results had to reflect that.

During the creation of the XLDBTumba01 run and while analysing our results together with the CLEF relevant judgements, we realized that in many cases, a simple query couldn't retrieve all the relevant documents. Take for instance, topic #204, for retrieving documents concerning avalanche victims. In the Portuguese Monolingual task, this topic had 7 relevant judgements, which contained the following relevant words of the 'avalanche' noun and the 'morrer' verb (*to die*) / 'morte' (*death*) family (Table 3):

We can see that it's impossible to achieve a good recall value with a query containing 'avalanche' 'morte' terms only. This is a situation that is not uncommon and IR systems must be able to deal with it.

We intend to extend our Web search system to provide much better results in situations where the documents are not rich in HTML features, such as hyperlinks and meta-tags. Tumba! is effective in named-page finding tasks, in particular when these have properly chosen titles and have multiple links.

Word	Rel #1	Rel #2	Rel #3	Rel #4	Rel #5	Rel #6	Rel #7
avalanche	x	x	x				
avalanches	x		x				x
avalancha				x	x	x	
mortos	x		x	x	x		
mortas							x
morte	x		x		x	x	x
morreu						x	
morreram		x	x				
morrido		x					
mata		x					

Table 3: Relevant words in the relevant documents of the topic 204

7 Acknowledgements

We would like to thank to Bruno Martins and Daniel Gomes for minor patches on tumba! components for the CLEF task and valuable comments, Marcírio Chaves and Lauro Nakayama for their manual retrieval and judgement of the documents for the XLDBTumba01 run, and Diana Santos and Luís Costa for their valuable suggestions and comments.

This work was financed by the Portuguese Fundação para a Ciência e Tecnologia through grant POSI / PLP / 43931 / 2001 (Linguateca) and by grant POSI / SRI / 40193 / 2001 (XMLBase Project).

References

[1] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan. *Searching the Web*, volume 1, pages 2–43. August 2001. `http://www.acm.org/pubs/contents/journals/toit`.

[2] Martin Braschler and Carol Peters. CLEF 2002 Methodology and Metrics, Advances in Cross-Language Information Retrieval: Results of the CLEF 2002 Evaluation Campaign. *Lecture Notes in Computer Science*, 2758, Spring 2003.

[3] Linguateca Centro de Recursos Distribuído para a Língua Portuguesa. `http://www.linguateca.pt`.

[4] Miguel Costa and Mário J. Silva. Sidra: a Flexible Distributed Indexing and Ranking Architecture for Web Search. In *Proceedings of the VIII Conference on Software Engineering and Databases JISBD 2003*, Alicante, Spain, November 2003.

[5] F. Couto, B. Martins, M. Silva, and P. Coutinho. Classifying Biomedical Articles using Web Resources : application to KDD Cup 02. DI/FCUL TR 03–24, Department of Informatics, University of Lisbon, July 2003.

[6] Francisco Couto, Mário Silva, and P. Coutinho. FiGO: Finding Genomic Ontology Terms in Text using Information Content. Granada, Spain, March 2004. BMC Bioinformatics Journal (accepted for publication).

[7] Pólo XLDB da Linguateca. `http://xldb.di.fc.ul.pt/linguateca/`.

[8] Tumba! Portuguese Web Search Engine. `http://www.tumba.pt`.

[9] Daniel Gomes, João P. Campos, and Mário J. Silva. Versus: a web repository. In *WDAS - Workshop on Distributed Data and Structures 2002*, Paris, France, March 2002.

[10] Daniel Gomes and Mário J. Silva. Tarântula - sistema de recolha de documentos da Web. In *CRC'01 - 4^a conferência de Redes de Computadores*, Novembro 2001.

[11] Notes on TREC Eval. `http://ir.iit.edu/~dagr/cs529/files/project_files/trec_eval_desc.htm`.

[12] Carol Peters and Martin Braschler. Cross-Language Evaluation Forum: Objectives, Results, Achievements. *Information Retrieval*, 7(1/2):7–31, January/April 2004.

[13] Público. `http://www.publico.pt`.

[14] Diana Santos and Paulo Rocha. CHAVE: topics and questions on the Portuguese participation in CLEF. This volume, 2004.

[15] Mário J. Silva. The Case for a Portuguese Web Search Engine. In *IADIS WWW/Internet 2003 Conference*, Novembro 2003.