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

Abstract. This paper describes the participation of the XLDB Group in the CLEF monolingual ad hoc task for Portuguese. We present tumba!, a Portuguese search engine and describe its architecture and the underlying assumptions. We discuss the way we used tumba! in CLEF, providing details on our runs and our experiments with ranking algorithms.

1 Introduction

In 2004, for the first time, CLEF included Portuguese document collections for monolingual & bilingual ad hoc 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 [7]. This year, the XLDB Group, from the University of Lisbon, made its debut 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 our official runs with the algorithms implemented for CLEF 2004, and Section 5 presents our results. Section 6 summarizes the conclusions we drew from this first participation in CLEF.

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 study 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.

Since January 2004, the XLDB Group hosts a node of Linguateca, a distributed language resource center for Portuguese [6].

The participation of the XLDB Group in the monolingual task for Portuguese with the tumba! search engine was motivated by two main reasons:

C. Peters et al. (Eds.): CLEF 2004, LNCS 3491, pp. 245–252, 2005.

1. Although we had previous experiences in evaluation contests, namely in the bio-text task of the KDD Cup 02 [4] and in the BioCreative workshop [5], this was our first opportunity to evaluate 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 tumba! in the Monolingual Task

3.1 Overview of tumba!

The tumba! search engine has been specifically designed to archive and provide search services to a Web community formed by those interested in subjects related to Portugal and the Portuguese people [8]. tumba! has been offered as a public service since November 2002.

tumba is mainly written in Java and built on open-source software: 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.

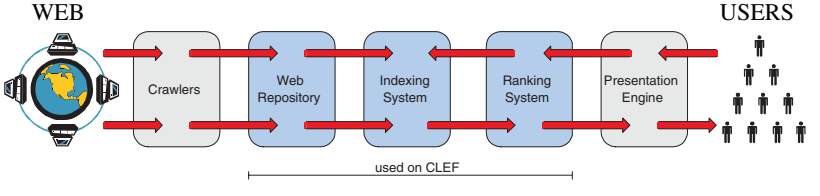

Fig. 1. tumba's architecture

The architecture of tumba! is similar to that of 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, and use these 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].