

A Penalisation-Based Ranking Approach for the Mixed Monolingual Task of WebCLEF 2006*

David Pinto^{1,2}, Paolo Rosso¹, and Ernesto Jiménez³

¹ Department of Information Systems and Computation, UPV, Spain

² Faculty of Computer Science, BUAP, Mexico

³ School of Applied Computer Science, UPV, Spain

{dpinto, prosson}@dsic.upv.es, erjica@ei.upv.es

Abstract. This paper presents an approach of a cross-lingual information retrieval which uses a ranking method based on a penalisation version of the Jaccard formula. The obtained results after the submission of a set of runs to the WebCLEF 2006 have shown that this simple ranking formula may be used in a cross-lingual environment. A comparison with runs submitted by other teams ranks us in a third place by using all the topics. A fourth place is obtained with our best overall results by using only the new topic set, and a second place was got by using only the automatic topics of the new topic set. An exact comparison with the rest of the participants is in fact difficult to obtain and, therefore, we consider that further detailed analysis of the components should be done in order to determine the best components of the proposed system.

Keywords: Retrieval models, Mixed-Monolingual search process.

1 Introduction

The current commercial search engines, such as Google and Yahoo, provide only monolingual information retrieval and, therefore, forums dedicated to the analysis and evaluation of information retrieval systems in a cross-lingual environment, such as WebCLEF [1], are needed. Since its first edition in 2005, the WebCLEF concern has been to deal with the EuroGOV corpus, which consists in a crawl of European governmental sites from approximately 27 different Internet domains [5,3]. The aim of this work is to evaluate a new similarity measure in the Mixed-Monolingual task of WebCLEF evaluation forum. The introduced formula is a variation of the Jaccard coefficient with a penalisation factor.

In the next section we will describe the way we have processed this corpus in order to obtain the index terms. Section 3 explains the model we have implemented, whereas its evaluation is presented in Section 4. Finally, a discussion of our participation and the obtained results in this evaluation forum are given.

* This work was partially supported by the MCyT TIN2006-15265-C06-04 project, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant.

2 Dataset Preprocessing

We have written two scripts in order to obtain the index terms of the EuroGOV corpus. The first script uses regular expressions for excluding all the information which is enclosed by the characters < and >. This script obtains very good results, but it is very slow and, therefore, we decided to use it only with three domains of the EuroGOV collection, namely Spanish (ES), French (FR), and German (DE). We have wrote another script based on the *html* syntax for obtaining all the terms considered valuable for indexing, which speeded up our indexing process. Unfortunately, major web pages do not strictly observe the *html* syntax and, therefore, we missed important information from those documents. We have implemented different methods for detecting the charset codification of those webpages that are not in UTF-8. We have observed that the charset codification detection is one the most difficult problem in the preprocessing step. Finally, we eliminated stop words for each language (except Greek) and punctuation symbols. The same process was applied to the queries.

3 The Penalisation-Based Ranking Approach

Nowadays, different information retrieval models are reported in literature [4] [2]. In this work, we have used a variation of the boolean model with ranking based in the Jaccard similarity formula. We named this variation “Jaccard with penalisation”, because it punishes the ranking score taking into account the number of terms that a query Q_i really matches when it is compared with a document D_j of the collection. The formula used is presented as follows:

$$Score(Q_i, D_j) = \frac{|D_j| \cap |Q_i|}{|D_j| \cup |Q_i|} - \left(1 - \frac{|D_j| \cap |Q_i|}{|Q_i|} \right) \quad (1)$$

As can be seen, the first component of this formula is the typical Jaccard approximation. The evaluation of this formula is quite fast, and allows its implementation in real situations. The obtained results by using this approach are presented in the next section.

4 Experimental Results

We have submitted three different runs in order to experiment with the use of diacritics in the corpus preprocessing step. We have renamed all the runs in this paper (cursive), with respect to the names reported in [1] (bold face) as follows: *WithoutDiac* (**ERFinal**), *WithDiac* (**ERConDiac**), *CDWithoutDiac* (**DPSinDiac**).

Table 1 shows the results obtained with each of the three different approximations submitted. The *WithoutDiac* run eliminates all diacritics in both, the