UPV/BUAP Participation in WebCLEF 2006*

[(1,2)]David Pinto, [1]Paolo Rosso & [3]Ernesto Jiménez
[1]Department of Information Systems and Computation,
Polytechnic University of Valencia (UPV), Spain
[2]Faculty of Computer Science,
B. Autonomous University of Puebla (BUAP), Mexico
[3]School of Applied Computer Science, UPV, Spain
{dpinto, prosso}@dsic.upv.es, erjica@ei.upv.es

Abstract

After our first participation in the Bilingual task of WebCLEF 2005, we have emigrated to a more challenging task. In this report we are presenting the results obtained after evaluating a set of topics in the Mixed-Monolingual task of WebCLEF 2006. Our efforts were focused on the preprocessing of the EuroGOV corpus which is itself a very challenging task, due to the high variety of errors that must be treated in order to correctly interpret the content of each document to index. Moreover, we have tested a new formula for the ranking of the documents retrieved, which is based on the Jaccard formula but includes a penalization factor. Results are low but encourage to investigate whether they are the result of a bad preprocessing process and/or the malfunction of the search engine components.

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

Measurement, Performance, Experimentation

Keywords

Retrieval models, Mixed-Monolingual search process

1 Introduction

Learning to deal with the high volume of data of Internet becomes more a need than a curiosity. Currently, we are witnesses of a big explosion of information available that must be adequately cataloged. Moreover, this information comes from all parts of the world, from very different cultures and different languages which makes this task even more difficult. At the moment, it would seem that only search engines such as Google and Yahoo could have enough resources for this challenge, but new proposals would provide advances from the scientific instead of the comercial

*This work was partially supported by the R2D2 (CICYT TIC2003-07158-C04-03) and ICT EU-India (ALA/95/23/2003/077-054) research projects, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant.

viewpoint. Certainly, forums dedicated to the analysis of information search and retrieval, more particularly in a cross-language environment, are needed.

The WebCLEF concern is about the evaluation of information retrieval systems using cross-lingual web pages. The justification of the WebCLEF track is based on the fact that many issues for which people turn to the web are essensially multilingual. In 2005 [6], the first edition of this competition was done in the framework of the Cross Language Evaluation Forum (CLEF) [4]. At that time, three tasks were proposed: mixed monolingual, multilingual, and bilingual English-Spanish. Currently, only one of those tasks was suggested: mixed monolingual; the reason of this action is derived from the very bad results obtained in the first edition in the multilingual task compared with those of the mixed-monolingual task. Due to this reason, the WebCLEF forum decided to focus this year on making robust the results in the mixed monolingual task, and later to make efforts for improving the translation resources that needed to be applied in the other task.

In 2005, we participated in the bilingual English to Spanish task [2]. In 2006 Mixed-Monolingual task, we have experimented using the EuroGOV corpus, which was compiled in 2005 before the WebCLEF campaign. This corpus consists in a crawl of governmental sites in Europe from approximately 27 differents Internet domains. A better description of this corpus can be found in [5]. Therefore, we will not describe the corpus, but the way we processed it in order to obtain the terms to index. The next section explains all the preprocessing steps we carried out. Section 3 explains the model implemented. In Section 4 we present our results and finally a discussion is given.

2 Preparing the data

The preprocessing phase of the EuroGOV corpus presents a big challenge, due to the written variants a government web page could have. We have found that a big amount of documents do not present a strict *html* syntax. We have written two scripts for obtaining the terms to be indexed from each document. The first script uses regular expressions for excluding all the information which is enclosed by the characters $<$ and $>$. Although this script obtains very good results, it is very slow and therefore we decided to used it only with three domains of the EuroGOV collection, namely Spanish (ES), French (FR), and German (DE).

On the other hand, we wrote a script based in the *html* syntax for obtaining all the terms considered interesting for indexing, i.e., those different than script codes (javascript, vbscript, style cascade sheet, etc), *html* codes, etc. This script speeded up our indexing process but it did not took into account that some web pages are incorrectly written and, therefore, we missed important information from those documents.

Another preprocessing problem consists in the charset codification, which leads to a even more difficult analysis. Although the EuroGOV corpus is given in UTF-8, the documents that made up this corpus does not neccesarily keep this charset. We have seen that for some domains, the charset codification is given in the *html* metadata tag, but also we found that this codification could be wrong, perhaps because it was filled without the supervision of the creator of that page, who may be does not know anything, and evenmore does not matter about charsets codifications. We consider it as the most difficult problem in the preprocessing process.

As usual in the information retrieval systems, we eliminated stopwords for each language (except Greek). A good repository of resources for this step is suministered by Jacques Savoy from the Institut interfacultaire d'informatique (see [1]). A variation on the elimination of diacritics was done; we discuss in detail this approach in Section 4. The same process was applied to the queries. The next section explains the model used in our runs.

3 Description of our model

Nowadays, different information retrieval models are reported in literature [3]. Perhaps the most popular model is the vector space model which uses the well-known $tf - idf$ formula, however, in

practice this model is not viable. We have used a variation of the boolean model with ranking based in the Jaccard similarity formula. We named this variation Jaccard with penalization, because it takes 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:

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

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 results obtained by using this approach are presented in the next section.

4 Results

At the moment of writing this paper, individual results are only known by each team in the competition, and therefore a comparative table with the other teams results is not presented. In Table 1 we show the results obtained with our approximation. We evaluated different runs, varying the use of diacritics and the preprocessing of the corpus.

The *WithoutDiac* run eliminates all diacritics in both, the corpus and the topics, whereas the *WithDiac* run only supresses the diacritics in the corpus. We can observe a expected reduction of Mean Reciprocal Rank (MRR), but it is not significatively high with respect to the first run. This is clearly derived from the amount of diacritics introduced in the topics of evaluation, which is not very high. An analysis of the queries in real situations may be interesting in order to determine whether the topics set is realistic. The last run (*CDWithoutDiac*) eliminates diacritization in both, the topics and corpus, but also tries a charset detection for each document to be indexed. Unfortunately, from the table we can observe that we did not success in our attempt.

Table 1: Evaluation of each run submitted

	Average Success at					
Run	**1**	**5**	**10**	**20**	**50**	**MRR over 1939**
WithoutDiac	0,0665	0,1423	0,1769	0,2192	0,2625	**0,1021**
WithDiac	0,0665	0,1372	0,1717	0,2130	0,2568	0,1006
CDWithoutDiac	0,0665	0,1310	0,1681	0,1996	0,2470	0,0982

We were short of time for finishing the preprocessing phase and, therefore, we indexed only 20 domains (we did not indexed the following domains: EU, RU, FI, PL, SE, CZ, LT). Due to this fact, only 1470 from 1939 topics were evaluated, which is approximately a 75,81% of the total of topics. Although we presented the MRR over 1939 topics, 469 topics related with the not indexed domains were not evaluated. Therefore, we are planning to continue evaluating the other queries and know the behaviour of our system in these domains.

The next section discusses findings in our first participation in the mixed monolingual task of WebCLEF.

5 Discussion

We have proposed a new approach for the ranking formula in a information retrieval system based on the Jaccard formula, but with a penalization factor. After evaluating this approach in the approximately 75% of queries from the WebCLEF competition, we obtained low results. We observed that our major problem was related to the preprocessing phase. The charset decodification must be improved in order to correctly interpret the data.

An evaluation of the use of diacritization in the task has shown that results are not significatively different, which may be suggesting that the set of queries provided for the evaluation does

not have a high number of diacritics. More investigation would determine whether this behaviour is realistic or must be tuned in further evaluations.

Even if a comparison with other results of the competition is still pending in order to determine how low are the results we obtained, we assume that our results are not very good from observing those from the last competition.

References

[1] Jacques Savoy: *Information about multilingual retrieval*, http://www.unine.ch/info/clef/.

[2] D. Pinto, H. Jiménez-Salazar, and P. Rosso: *BUAP-UPV TPIRS: A System for Document Indexing Reduction on WebCLEF*, Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Lecture Notes in Computer Science, Vol. 4022, Springer-Verlang, 2006.

[3] G. Salton: *Automatic Text Processing*, Addison-Wesley, 1989.

[4] J. Savoy, P. Y. Berger: *Report on CLEF-2005 evaluation campaign: Monolingual, bilingual, and GIRT information retrieval*. In C. Peters, Clough, P., Gonzalo, J., Jones, G., Kluck, M., Magnini, B. (Ed.), Working notes of CLEF 2005, 2005.

[5] B. Sigurbjörnsson, J. Kamps, and M. de Rijke: *EuroGOV: Engineering a Multilingual Web Corpus*, Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Lecture Notes in Computer Science, Vol. 4022, Springer-Verlang, 2006.

[6] B. Sigurbjörnsson, J. Kamps, and M. de Rijke: *WebCLEF 2005: Cross-Lingual Web Retrieval*, Accessing Multilingual Information Repositories: 6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Lecture Notes in Computer Science, Vol. 4022, Springer-Verlang, 2006.