A Penalisation-Based Ranking Approach for the Mixed Monolingual task of 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

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 rank 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.

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

In the last years we have been witnesses of a big explosion of information available in Internet. The correct classification of the mentioned information is the most important challenge for the information retrieval field. The fact that the information we are dealing with comes from all around the world and, therefore, from very different cultures with different languages, makes this task even more difficult. Moreover, the current commercial search engines, such as Google and Yahoo, provide only monolingual information retrieval, that is, given a query in a specific language, those systems retrieve query related documents which are written in the same language. In other words, current search engines do not consider the query language when these keywords are matched against the target document set.

Forums dedicated to the analysis of information search and retrieval, more particularly in a cross-language environment, are then 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 in essence multilingual. In 2005, the first edition of this competition was done in the framework of the Cross Language Evaluation Forum (CLEF) [1]. In this edition, we have participated in the mixed-monolingual task of the WebCLEF 2006 by 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] and, therefore, in the next section we will not describe the corpus, but the way we processed it 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 competition are given.

2 Dataset Preprocessing

The EuroGOV corpus preprocessing phase has presented a big challenge, due to the written different variants of the government web pages. We have found that a big amount of documents do not present a strict *html* syntax and, therefore, we have written two scripts for obtaining the index terms of each document. 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 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. This script speeded up our indexing process but it did not took into account that some web pages do not strictly observe the *html* syntax; and, therefore, we missed important information from those documents.

Although the EuroGOV corpus is given in UTF-8, the documents that made up this corpus do not neccesarily keep this codification. We have seen that for some domains, the charset codification is given in the *html* metadata tag, but also we have found that very often this codification is wrong. We consider the charset codification detection the most difficult problem in the preprocessing step.

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

3 The Penalisation-Based Ranking Approach

Nowadays, different information retrieval models are reported in literature [4] [2]. The most popular is the vector space model, however, in practice this model is not viable. 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:

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

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.

[1] http://www.unine.ch/info/clef/

4 Experimental Results

We have submitted three different runs to the WebCLEF 2006 competition in order to experiment with the use of diacritics and the corpus preprocessing. We have renamed all the runs in this paper with respect to the names reported in [1]. Their new names as well as their name referred in that paper are enumerated with cursive (emphasize) and bold face, respectively, 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 corpus and the topics, whereas the *WithDiac* run only supresses the diacritics in the corpus. We may observe an expected reduction of the 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 evaluation topics set, 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 tried 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					
Team	**Run**	**1**	**5**	**10**	**20**	**50**	**MRR over 1939**
rfia	*WithoutDiac*	0,0665	0,1423	0,1769	0,2192	0,2625	**0,1021**
rfia	*WithDiac*	0,0665	0,1372	0,1717	0,2130	0,2568	0,1006
rfia	*CDWithoutDiac*	0,0665	0,1310	0,1681	0,1996	0,2470	0,0982

Table 2 shows a summary of all the best participant runs submitted to the mixed monolingual task of WebCLEF 2006. The Mean Reciprocal Rank (MRR) scores are reported for both the original and the new topic set. The first column indicates the name that each team had in the competition, whereas the second column indicates the name of their best run. The scores shown in that table rank us in a third place.

Table 2: Best runs for each WebCLEF 2006 participant in the mixed monolingual task

Team Name	**Run Name**	**MRR for the original topic set**	**MRR for the new topic set**
isla	CombPhrase	0.2001	0.3464
hummingbird	humWC06dpcD	0.1380	0.2390
rfia	WithoutDiac (ERFinal)	0.1021	0.1768
depok	UI2DTF	0.0918	0.1589
ucm	webclef-run-all-2006	0.0870	0.1505
hildesheim	UHiBase	0.0795	0.1376
buap	allpt40bi	0.0157	0.0272
reina	USAL_mix_hp	0.0139	0.0241

In table 3(a) we can see the best overall results by using only the new topic set. Here we have obtained a fourth place, according to the average among the automatic and the manual topic scores. Whereas, in Table 3(b) we may observe the results by using only the new automatic generated topics. Our second place shows that the penalisation-based ranking is working well for the task proposed in this competition. Interesting is to see that our approach obtains a better behaviour on the new than the original topics. As can be seen in [1], the new topics were mostly automatically generated; whereas the original where all manually generated. Further investigation would analyse the above mentioned behaviour of the penalisation-based ranking approach presented in this paper.

Table 3: Best overall runs for each WebCLEF 2006 participant by using: *(a)* the new topic set, and *(b)* only the automatic generated topics

Team Name	automatic	manual	average
isla	0.3145	0.4411	0.3778
hummingbird	0.1396	0.5068	0.3232
depok	0.0923	0.3386	0.2154
rfia	0.1556	0.2431	0.1993
hildesheim	0.0685	0.3299	0.1992
ucm	0.1103	0.2591	0.1847
buap	0.0080	0.0790	0.0435
reina	0.0075	0.0689	0.0382

(a)

automatic	manual	average
0.3145	0.3114	0.3176
0.1396	0.1408	0.1384
0.0923	0.1024	0.0819
0.1556	0.1568	0.1544
0.0685	0.0640	0.0731
0.1103	0.1128	0.1077
0.0080	0.0061	0.0099
0.0075	0.0126	0.0022

(b)

5 Discussion

We have proposed a new approach for the ranking formula in an information retrieval system based on the Jaccard formula, but with a penalisation factor. After evaluating this approach in the approximately 75% of queries from the WebCLEF competition, we have obtained the third place in the overall results, among eight participant teams.

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. Further investigation would determine whether this behaviour is realistic or must be tuned in further evaluations.

The comparison by using only the new topic set ranks the proposed system in fourth and second place for the overall results which uses both, the new topic set and the automatic generated new topics, respectively. We consider that the system proposed may be improved by taking into account a better understanding of the preprocessing phase.

Acknowledgements

This work is a revised version of the paper "UPV/BUAP Participation in WebCLEF 2006", published in the abstract proceedings of the CLEF 2006. It was partially supported by the MCyT TIN2006-15265-C06-04 project, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant.

References

[1] K. Balog, L. Azzopardi, J. Kamps, M. d. Rijke: *Overview of WebCLEF 2006*, In proceedings of the CLEF'06: Cross-Language System Evaluation Campaign, A. Nardi, C. Peters and, J. L. Vicedo (Eds.), page 51, 2006.

[2] W. Kraaij, M. Simard, J.Y. Nie: *Embedding Web-based Statistical Translation Models in Cross-Language Information Retrieval*, Computational Linguistics, 29(3):381–419, 2003.

[3] D. Pinto, H. Jiménez-Salazar, and P. Rosso: *BUAP-UPV TPIRS: A System for Document Indexing Reduction on WebCLEF*, Accessing Multilingual Information Repositories: CLEF 2005, LNCS, Vol. 4022, Springer-Verlang, 2006.

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

[5] B. Sigurbjörnsson, J. Kamps, and M. de Rijke: *EuroGOV: Engineering a Multilingual Web Corpus*, Accessing Multilingual Information Repositories: CLEF 2005, LNCS, Vol. 4022, Springer-Verlang, 2006.