REINA at WebCLEF 2006. Mixing Fields to Improve Retrieval

Carlos G. Figuerola, José L. Alonso Berrocal, Ángel F. Zazo Rodríguez, Emilio Rodríguez
REINA Research Group, University of Salamanca
reina@usal.es

Abstract

This paper describes the participation of the REINA Research Group of the University of Salamanca at WebCLEF 2006. The task in that we have participated this year is the Monolingual Mixed Task in Spanish. To select web pages of the EuroGov collection in Spanish, the wide collection was processed with a language guesser, searching for pages in Spanish. All pages in the .es domain were also pre-selected. Our focus, this year, is to test pre-retrieval ways of mixing fields or elements of information in web pages, as well as to test the retrieval capacity of these fields. Mixing terms from several sources in a only index can be achieved, in retrieval systems based on the vector space model, operating on the term frequency in the document, if we use a $tf \times idf$ schema of weigthing. BODY field is, by the way, the most powerfull from the point of view of retrieval, but ANCHORS of backlinks add a considerable improvement. META fields, nevertheless, contribute little to the improvement in retrieval.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

web pages retrieval, information retrieval, web search, combining fields

1 Introduction

This article describes the participation of the REINA Research Group in the track WebCLEF 2006. The task in that we have participated this year is the Monolingual Mixed Task in Spanish. The document collection is the same one that in 2005 [12], as well as part of the topics.

Nevertheless, the last year we exclusively limited our work to documents or pages pertaining to the domain .es [3]. In this time, nevertheless, we have chosen to extend the document base to all those that, being in other domains, also are in Spanish.

To be within the domain .es, on the other hand, does not mean that the page is necessarily in Spanish; many pages are in some of the other languages spoken in Spain (Catalan, Euskera, Galician...), and others are versions internationalized in English, specially, but also in French and German. Of another side, there are pages in Spanish in several of the remaining European domains

that comprise of the collection used in WebCLEF. Thus, if really we want to look for all the pages that are in Spanish, is necessary to process all the collection and to select those pages that are in Spanish. This is what we have done, adding them the totality of the pages pertaining to the `.es` domain.

Unfortunately, the headers of the pages provide trustworthy information, neither on the language nor on other thing; in many cases, the headers simply are empty, whereas in others they have contents that do not correspond with the reality. Therefore, we have had to resort to a language detector to select the pages in Spanish within the diverse European domains. The selected detector is *TextCat* [8], a software based on the elaboration of patterns with n-grams for each possible language, and the later categorizacin of the text whose language is desired to guess [2].

The selection of the topics, of another side, is easy, since the language of these comes signalized clearly. All the topics in Spanish have been processed, and this includes those of type named pages, *Home Page*s, elaborated manually and the elaborated ones of automatic way. All has been dealed with the same way.

This paper is organized of the following way: in the next section the adopted approach to solve the task is described; in the following section the problems and options adopted in the lexical analysis and extraction of terms of the pages are exposed. Next, usable fields or elements of information in indexing and retrieval are discussed. In the following section we describe runs carried out, and the results are discussed. Finally, conclusions are provided.

2 Our Approach

The adopted basic strategy is: first to find the most relevant pages for each topic and, based on the type of topic, to rearrange the found more relevant documents list. Thus, besides to preselect pages which they will form the collection (those that belongs to the domain `.es` plus that they are in Spanish in other domains), our work has two basic parts: to find pages relevant and after to rerank those found pages. To find the relevant pages for a given topic can approach by means of a conventional system of indexing and retrieval, like, for example, the based ones on the model of the vectorial space [11]. Nevertheless, in a Web page there are more informative elements in addition to the text that the user sees in the window of his navigator. Even within that same text, we can find certain structures that can help, along with those other informative elements, to improve the retrieval.

For example, some of those mentioned elements are the field `TITLE`, some `META` tags, the `ANCHORS` of backlinks, etc. Within the field `body`, that is what it visualizes in the navigator, we can differentiate parts that use different typographies, for example. Thus, we have different sources of information that we must mix or fuse. Two basic strategies of fusion have been proposed: fusion pre-retrieval and fusion post-retrieval [1], [6]. This last strategy is the one that we applied in webCLEF 2005 and, basically, consists of building an index by each element of information to fuse. Whenever there is to solve a topic, it is executed against each one of the indexes and the results obtained in each index are merged later.

In this time we have wanted to test the strategy pre-retrieval, that consists of elaborating an only index with the terms of all the elements, but weighted in a different way. Once built that only index, the topics are executed normally against him. Naturally, in the elaboration of that only index we can wish more or less to value the originating terms of such-and-such field or element of information; this allows us to use a mixture that, if it is well in tune, would have to provide good results in the retrieval.

The application of a different weight from the different components from the mixture, in our case, is easy. Since for the indexing and retrieval we use our software *Karpanta* [4], based on the good well-known model of the vector space, we do not have more to operate on the frequency of each term in each one of the components of the mixture.

In this case we applied a scheme of weight `ATU` (slope=0.2) [13], but the idea is similar for any scheme of weight based on $tf \times idf$. We can to apply a coefficient to tf based on in what component of mixture appears term, so that it makes increase or diminish the weight of that term

based on in what component it appears, without letting consider the frequency in the document and the IDF.

As for the second part of our strategy, to rearrange the list of documents or pages retrieved based on the type of topic, we have only considered the case of the topics of type *Home Page*. For the location of *Home Pages*, a simple strategy was followed. First, the topics that they could be of type *Home Page* were manually detected, although this one is a strongly subjective valuation. Then, for the results of those topics, a coefficient was calculated on the basis of two criteria; first, a simple heuristic based in the existence in `TITLE` of certain expressions: *main*, *welcome page*, *etc.*. A heuristic similar was also applied to the name of the file (for example, *main.html*, *home.html*, *etc.*). The other criterion is the length of the URL, understood like the levels of path of the URL [7]: smaller number of levels makes more probable that it is a *Home Page* [10], [14].

The obtained coefficient simply is multiplied by the similarity of each page retrieved in each topic of type *Home Page*, rearranging the results of each retrieval.

2.1 Lexical Analysis and Extraction of Terms

The first operation, previous to any other, is the conversion of the Web pages to plain text. This is necessary even to be able to determine the language of each page and to select it or not. The obtaining of the plain text is not trivial and she is not free of problems. As it were already said before, one cannot trust that the standards are followed. Not even it is possible to be guaranteed that the content is HTML, although the page begins with the appropiate tag. So, one can be directly with binary code, PDFs and similars. With the tags `META` it happens the same; even though these are present, not always offer correct information.

In fact, many pages not even contain text; so, first is to determine the type of content. The old one and known command *file*, well in tune, can help to determine the real content of each page. In addition, it will inform to us into another important data: the used system of codification. For the pages in Spanish, this one usually is *ISO 8859* or *UTF-8*; it is necessary to know this information to treat the special characters suitably. The conversion to plain text, when the page turns out to be HTML, is carried out by means of *w3m*. There are other converters, but after several tests the best results were obtained with *w3m*.

Once obtained the plain text, we can determine the language with *TexCat*. Those documents in Spanish, as well as all pertaining to the domain `.es` will form our collection. Of these, it is precise to extract terms and to normalize them somehow. Basically, the characters are turned to small letters, the accents are removed, as well as numbers, orthographic and similar characters. The stop words are eliminated and the terms are passed through an enhanced s-stemming [5].

2.2 Used Fields

Are diverse the elements or sources of information that we can consider in a Web page. The base is the field `BODY`, obviously, but in addition we can use the field `TITLE`, that seem clearly descriptive, as well as diverse tags `META` that usually are used for these purposes, in special `META content="Description"` and `META content="Keywords"`. Nevertheless, as were already indicated in [3], these fields are not present of a uniform way in all the pages. Many do not contain them, and others contain `META` values of automatic form by the programs that have produced those pages Web.

In other cases, although the values are put manually by the authors of the pages, are little operative. Additionally, we can think that terms that appear in outstanding typography can be more representative. The labels or fields `H1`, `H2`, etc. are an example. Unfortunately, the use of these tags is not uniform and either withdraws in favour of the definition of sources and specifics sizes of characters, more difficult to process.

Another important element is the `ANCHORS` of backlinks that receive the pages. Of more or less brief form, these `ANCHORS` describe the page with which they connect; this description is important, because it is done by somebody different one from the author of the page that we want to index. We can be conceited that this description can add terms different from the used ones

RUN 1	BODY(fd=1)
RUN 2	BODY(fd=1)
	ANCHOR(fd=1)
	TITLE (fd=1.5)
	META-DESC (fd=1.5)
	META-TITLE (fd=1)
	META-KEY (fd=0.5)
	H1 (fd=0.8)
	H2 (fd=0.8)
RUN 3	Same as RUN 2 + *Home Pages* boosting

Table 1: Official Runs

	RUN 1	**RUN 2**	**RUN 3**
Average success at 1	0.0067	0.0098	0.0093
Average success at 5	0.0129	0.0175	0.0186
Average success at 10	0.0170	0.0201	0.0211
Average success at 20	0.0201	0.0248	0.0248
Average success at 50	0.0294	0.0315	0.0315
MRR	0.0100	0.0137	0.0139

Table 2: Results of the Official Evaluation

in the own page. Nevertheless, many of these ANCHORS can make reference to very concrete parts of the pointed page; also, there are pages with many backlinks and many ANCHORS, and others with few or no. And, in any case, we have the problem to obtain these ANCHORS. In our case, we have processed the totality of the EuroGov collection, to obtain all the ANCHORS and links towards pages in Spanish or pertaining to the domain .es. It is evident that, outside EuroGov will be more links towards those pages, but we do not have way of obtain them.

Finally, we have built indexess with the following elements: BODY, TITLE, meta-title, meta-description, meta-keywords, H1, H2, ANCHORS.

3 Runs Executed

With topics and estimations of relevance from WebCLEF 2005, we made a phase of training that allowed us to test several mixtures of elements in diverse proportions. Thus, we have made three runs officials and several unofficial ones. First run was executed against an only index that was built based on the BODY field, and it will serve to us as comparison base. In the second run, the index is a mixture of all the fields mentioned before in the proportions indicated in the table. Third one is based on the same index that second, but with boosting of *Home Pages* added, based in the size of URLs and the simple heuristic commented before.

The results of these officials runs confirm clearly the advantage of the use of those additional elements of information. Of them, it seems that most useful they are the ANCHORS of backlinks.

Of another side, there are several runs unofficials which confirm the official results. The graphic shows the results. Each run executes against an index elaborated with the terms of each field or element of information (fd=1). Each run works with one single of these fields, without using the terms that appear in BODY. The fields H1 and H2 are not in the graphic, because they produce extremely low results. It is evident that many documents or pages lack one or several of such fields, reason why never could be retrieved of this form; but it is a good way to separately verify the capacity of retrieval of these fields.

As it is possible to wait for, each field separately produces worse results than BODY, which is

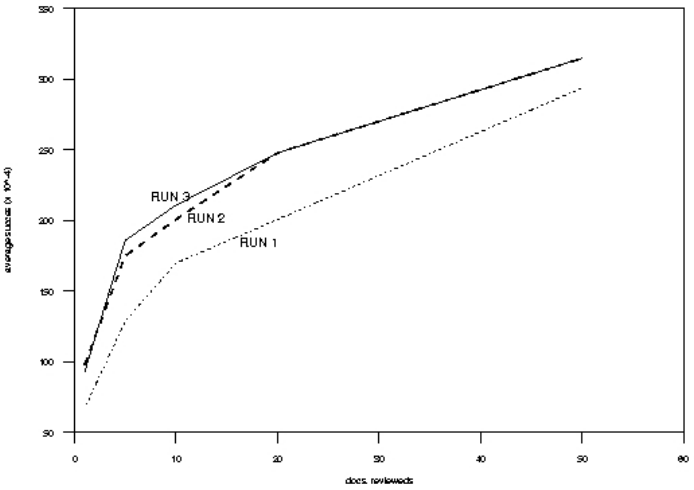

Figure 1: Results of the Official Runs

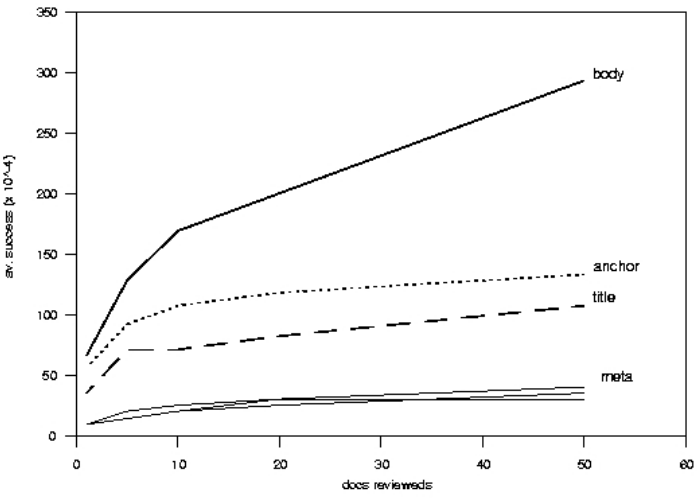

Figure 2: Unofficial Results by Individual Fields

normal, since in BODY it is where it is the visualizeable text of the page. But, aside from BODY, the greater field with being able of retrieval is the ANCHOR of backlinks, although in many cases those ANCHORS are very short. But it seems that, like saying is had, the descriptions that others do of a page are quite effective for their retrieval. It follows short distances to the field TITLE to him, something foreseeable. Nevertheless, the fields based on the META tags offer poor results, to enough distance of ANCHORS and TITLE, although they are fields oriented specifically to the retrieval. Although there is little difference between the results of the three META observed (title, description and keywords) is this one last one, peculiarly, the one that worse works of the three.

4 Conclusions

We have described our participation in WebCLEF 2006, the adopted strategy, the conducted experiments and the obtained results. The use of fields or additional elements of information to the text or BODY of the pages allows to improve the results of the retrieval. A form to use these fields is to elaborate an only index with the terms that appear in them, along with the words that appear in BODY of the page. This requires to weight the terms of each field in form different, adjustable on a empirical way.

Of all those fields, it seems that most effective is the formed one by the ANCHORS of backlinks that receives each page. The field TITLE also contributes of remarkable form to the improvement of the retrieval. The content of the META tags, nevertheless, seems of utility reduced, from the point of view of the retrieval.

Acknowledgements

This research has been partially funded by the Government of the Autonomus Community of Castilla y León as project ref. SA089/04.

References

[1] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, Ophir Frieder, and Nazli Goharian. On fusion of effective retrieval strategies in the same information retrieval system. *Journal of the American Society for Information Science and Technology (JASIST)*, 55(10):859–868, 2004.

[2] William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In *Third Annual Symposium on Document Analysis and Information Retrieval. April 11-13, 1994, Las Vegas, Nevada*, pages 161–175, 1994.

[3] Carlos G. Figuerola, José L. Alonso Berrocal, Ángel F. Zazo Rodríguez, and Emilio Rodríguez. REINA at the WebCLEF task: Combining evidences and link analysis. In Peters [9].

[4] Carlos G. Figuerola, José Luis A. Alonso Berrocal, Ángel F. Zazo Rodríguez, and Emilio Rodríguez Vázquez de Aldana. Herramientas para la investigación en recuperación de información: Karpanta, un motor de búsqueda experimental. *Scire*, 10(2):51–62, 2004.

[5] Carlos G. Figuerola, Ángel F. Zazo, Emilio Rodríguez Vázquez de Aldana, and José Luis Alonso Berrocal. La recuperación de información en español y la normalización de términos. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 8(22):135–145, 2004.

[6] Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*. NIST Special Publication 500-215, 1993.

[7] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 27–34. ACM Press, 2002.

[8] Gertjan van Noord. Texcat language guesser. *http://www.let.rug.nl/∼ vannoord/TextCat*

[9] Carol Peters, editor. *Results of the CLEF 2005 Cross-Language System Evaluation Campaign. Working notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria*, 2005.

[10] Vassilis Plachouras, Fidel Cacheda, Iadh Ounis, and Cornelis Joost van Rijsbergen. University of Glasgow at the Web Track: Dynamic application of hyperlink analysis using the query scope. In *The Twelfth Text REtrieval Conference (TREC 2003)*. NIST Special Publication 500-255, 2003.

[11] Gerard Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communication of the ACM*, 18:613–620, 1975.

[12] Borkur Sigurbjornsson, Jaap Kamps, and Maarten de Rijke. Overview of WebCLEF 2005. In Peters [9].

[13] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 18–22, 1996, Zurich, Switzerland (Special Issue of the SIGIR Forum)*, pages 21–29. ACM, 1996.

[14] Stephen Tomlinson. Robust, web ad terabyte retrieval with Hummingbird Searchserver at TREC 2004. In *The Thirteen Text REtrieval Conference (TREC 2002)*. NIST Special Publication 500-261, 2004.