Web Page Retrieval by Combining Evidence*

Carlos G. Figuerola, José L. Alonso Berrocal,
Angel F. Zazo, and Emilio Rodríguez Vázquez de Aldana

University of Salamanca, REINA Research Group,
reina@usal.es
http://reina.usal.es

Abstract. The participation of the REINA Research Group in WebCLEF 2005 focused in the monolingual mixed task. Queries or topics are of two types: *named* and *home pages*. For both, we first perform a search by thematic contents; for the same query, we do a search in several elements of information from every page (title, some meta tags, anchor text) and then we combine the results. For queries about *home pages*, we try to detect using a method based in some keywords and their patterns of use. After, a re-rank of the results of the thematic contents retrieval is performed, based on Page-Rank and Centrality coeficients.

1 Introduction

Our participation in WebCLEF 2005 focused on the monolingual mixed task in Spanish. The task has a two-fold objective: to find *named web pages* and *home pages*. Each query has a single valid response and both types of queries are mixed and we do not know a priori to which type of query each one pertains.

In principle, the basic approach consists of finding pages whose content is relevant to each query; the valid response is expected to be found among the first pages retrieved and a better or worse positioning depends on the techniques applied in the search.

In cases of queries searching for a *home page* we apply a procedure that reorders the list of documents retrieved, taking into account, besides their similarity to the query, different types of evidence that point to their being *home pages*. A further problem is that we do not know a priori which queries are searching for *home pages* and which are not, so we must include a procedure to analyze the queries and determine which ones are searching for *home pages* and which are not.

The paper is organized as follows: in section 2 we offer a description of the part of the document collection that we worked with; section 3 describes the approach applied; the section following that reports on the experiments carried out and their results and finally, the last section gives our conclusions.

* This research has been partially supported by a grant from the Junta de Castilla y León, project number SA089/04.

2 The Document Collection

Our participation was limited to the `.es` domain. It has a total of 35,168 documents; not all the pages are `HTML` and it is not always easy to identify the document format; the `Content-type` is empty in many of the documents. For this year, the queries were limited to `HTML` documents and the organizers facilitated a blacklist of 4,365 documents that are not in `HTML`.

Table 1. Blacklist for `.es` domain

Format	Number of docs.
PDF	4040
MS Word	315
empty docs	6

However, there are documents in other formats that are not on the blacklist. Thus, of the 35,168 documents in the `.es` domain, 8642 are not labeled `<HTML>`.

Furthermore, the documents seem to have been truncated to a size of approximately 64K, and in the binary files, such as PDF files, the characters `chr(0)` seem to have been replaced by `chr(32)`.

2.1 Topics

There are 118 topics in Spanish, 59 searching for *home pages* and 59 for *named pages*. The concept of *home page*, however, is fuzzy; the consideration of some of the searched pages as *home* is quite debatable.

In addition, there are some mistakes in the topics set. Thus, some topics are duplicated, or even triplicated. Some of them, with different correct page as answer in the *qrels* file. Some topics are a formulation too wide. By example, topic `WC0098`: *Consejería de Educación y Cultura*; there are, in Spain, 17 Autonomous Communities and every one of them has a Council of Education and Culture. Besides, we have found that many embassies have also a *Consejería de Educación y Cultura*, and there is a lot of embassies. How can a search engine determine which of them is the right answer?

A few topics have as correct answer a page which is not in the `.es` domain. This is, maybe, right; but, since we work only in the .es domain, we cannot find the correct page anyway.

3 Our Approach

As mentioned earlier, the basic idea was to find pages or documents closest to each query, and, in the case of *home page* type queries, prioritize on the list of retrieved documents the pages most likely to be *home pages*. This also obliged us to analyze the queries to determine their type.

The first part of our task, to find the pages most similar to each query, could have been approached using a classic scheme for document retrieval. However, web pages contain informative elements other than the text seen in the windows of navigators. We could thus use these elements to refine the retrieval.

3.1 Combining Evidence

The list of elements that we can take into account in web pages is long, but we focused on the following:

- The body field, which seems to be the most important
- The title field
- The contents of some META tags, as in the case of Description and Keywords
- The anchor text of incoming hyperlinks to a page.

All these elements supply evidence that we can somehow combine to find the pages most similar to each query. There are several ways to make this fusion or combination, and the first choice is whether to do the fusion before making the query or after. We opted to do it afterwards, and therefore the procedure applied was the following:

- build an index with the terms of each of the elements to be taken into account
- execute the query in each of these indexes.
- fuse the results obtained with each of the indexes

For the first step we used our *Karpanta* software [1], based on the well-known vector model, and built indexes of the fields `BODY`, `TITLE`, `META Description`, `META keywords`, and anchor text. The weights of the terms were calculated according to the classical scheme based on $tf \times IDF$ known as `atc`. In all cases the empty words were previously eliminated, applying a list of some 300 words in Spanish; also, an improved s-stemmer [2] was applied.

The sizes of the resulting indexes were uneven, as were the fields or elements on which the indexes were based. Almost all the HTML pages contained a `BODY` field (some only have *java* scripts and the like), but this was not the case for the rest of the indexes. So, 71.5 % of the pages in the `.es` domain contained a `TITLE` field and the mean length of these titles was 40 characters, which means they are very short titles.

The `META Description` tag or field was only present in 16.9 % of the documents, with a mean size of 38.6 characters. Of these documents, in 7.4 % of the cases the `META Description` coincided exactly with that of the `TITLE` field. The keywords (`META Keywords` field) only appeared in 24.7 % of the documents, with a mean of 7.7 words per document. As regards backlinks, 24.7 % of the documents had none (from inside the collection), and those that did receive them did so with a mean of 9 backlinks per document. The text of these backlinks, on the other hand, was very short (18.7 characters), although perhaps very significant. It thus seems clear that, except for the body field, the rest of the elements are limited in importance, since they were not present in large amounts of documents. For the fusion or combination of the resulting lists in each of the retrievals

on each index, first the coefficients of similarity were normalized based on the *z-score* [3] and then the normalized lists were fused using the CombMNZ algorithm [4], modified in order to be able to weight differently the results obtained with each index:

$$Score = \sum_{i=1}^{n} score_i \times k_i \times (number\ of\ score\ != 0) \quad (1)$$

There are other fusion procedures that can be applied [4,5,6,7]. Most of them are based on the combination of the coefficients of similarity obtained after executing the query in each index, but it is also possible to work with the positions in the lists of documents retrieved in each index [8]; this algorithm is attractive because of its simplicity, since it is not even necessary to previously normalize the scores or coefficients.

3.2 Finding *home pages*

The first step was to determine which queries are searching for *home pages*. The concept of *home page*, however, is diffuse, and therefore not everyone would consider as *home pages* some of the correct answers to some queries.

In an exploratory phase, different *home pages* of the `.es` domain were examined manually, particularly the `TITLE` field, with the idea that a query that hoped to find that page was probably quite similar to its title. Likewise, the *home page* type queries used in TREC were also consulted manually. They are in English, but once translated can give an idea of the structure and characteristics of this type of query.

During this phase some common elements were found in the structure of the home page queries. This structure has a lot to do with the use of specific terms related to the *home page* being looked for. Thus, pages of this type are those that give entry to the webs of certain institutions: ministries, institutes, schools, etc., and, as a consequence, these words will appear in the query [3].

Furthermore, they appear in certain positions and accompanied, before and after, by certain auxiliary words (articles and other connectors). This allowed us to build a series of patterns of *home page* queries to which a simple heuristic was added: the appearance of expressions such as *home page*, *portal*, etc. Once these supposedly *home page* queries had been identified using this system, the results of a search resolved by means of a combination of evidence such as those seen in the section above were reordered so as to place at the top those pages which, being relevant in the contents, were most likely to be *home pages*.

To determine which of the pages found can be *home pages*, several techniques have been described which are non-exclusive and can be combined with each other. The most well-known techniques use two types of information: the `URL` structure of the page, on the one hand, and links analysis, on the other.

The techniques based on the `URL` structure operate with the depth of that structure. Kraaij, Westerveld and Hiemstra [9] studied the statistical distribution of home pages in the different depth levels of the `URL`, as well as Beitzel and

colleagues [3]. Plachouras, Ounis, Rijsbergen and Cacheda [10] also used criteria based on the length of the URL, as did Tomlinson [11].

The techniques based on the analysis of links have also been widely used. Although judged to be of less usefulness in searches by content, they seem to be effective in recognizing *home pages* [12]. Different coefficients have been used, ranging from simple *in* and *out-degrees* [13] to *page-rank* [14] or *HITS* [15]. We tried with *page-rank* [16] and with the *centrality* index [17], both based on backlinks.

4 Experiments Performed

We performed official and unofficial experiments. Our aim was to determine what elements or evidence would be useful in the search for contents and what indexes based on links analysis seemed to be more effective in finding *home pages*.

The official results are shown in Table 2. USAL0 was used as a baseline for comparison and this was carried out with the queries in Spanish on pages in the .es domain. Only the BODY field of the pages was indexed, and all the queries were processed in the same way.

USAL1 combines results from the BODY, META Description fields and the text of the backlinks to each page.

USAL2 adds META Keywords to the fields of USAL1. USAL3 and USAL4 attempt to apply specific methods to locate home pages. From the results of USAL1 an attempt was made to detect *home page* type queries, and the results of these queries were re-ordered with *Page-Rank* in USAL3 and with *Centrality* in USAL4.

4.1 Evaluation

Table 2 shows the results of the official evaluation of the experiments. However, we have seen before some problems about the queries (duplicated ones, right answers in anothers domains). So, we have carried out an unofficial evaluation, removing erroneous topics: duplicated ones (even triplicated), right answers out of the .es domain, badly formulated queries. Classification in *home* and *named pages*, although debatable, we have left it as it was.

Table 2. Results (.es domain only) of the Official Evaluation

	USAL0	**USAL1**	**USAL2**	**USAL3**	**USAL4**
success at 1	0.1343	0.1642	0.1567	0.1940	0.1567
success at 5	0.3134	0.4254	0.3657	0.4776	0.4179
success at 10	0.3731	0.5000	0.4776	0.5522	0.4925
success at 20	0.3955	0.5970	0.5821	0.6493	0.6269
success at 50	0.6269	0.7463	0.7090	0.7537	0.7313
MRR	0.2193	0.2796	0.2553	0.3214	0.2776

Table 3. Unofficial Evaluation (`.es` domain only)

	USAL0	USAL1	USAL2	USAL3	USAL4
success at 1	0.1622	0.1982	0.1892	0.2162	0.1892
success at 5	0.3694	0.5135	0.4414	0.5586	0.5045
success at 10	0.4324	0.6036	0.5676	0.6486	0.5946
success at 20	0.4595	0.6847	0.6667	0.7207	0.7117
success at 50	0.7117	0.8378	0.7928	0.8468	0.8378
MRR	0.2611	0.3339	0.3045	0.3667	0.3255

Table 4. Most frequent keywords in `.es` domain

Keyword	times
cultura	1864
ministerio	1624
investigacion	1202
spain	1174
administracion	1171
politica	1169
informacion	1169
policy	1168
ministry	1168
research	1168
telecommunications	1168
information	1157
espaa	1157
industria	1126
turismo	1119
comercio	1080
energia	1012
telecomunicaciones	990
industry	962
trade	962
commerce	962
energy	962
tourism	962
parques nacionales	658

4.2 Results

It seems clear that working with more elements than just the `BODY` field improves retrieval; this seems to be true for `TITLE`, `META Description` and anchor text. However, the use of `META Keywords` made the results worse. This may seem surprising (certain simple retrieval systems are based on this field alone), but if we examine the use that the different pages make of it we see that, at the least, it is a strange use. Table 3 shows the keywords expressions (not individual terms) most used in the `.es` part of the collection.

For the most part these are very generic terms, not very useful for searches made in a government collection. Many of them are included in pages also translated into English, and some of them directly in English, without their Spanish counterpart (even though the rest of the page is in Spanish).

A manual examination of some of the pages of the collection showed that there are pages (particularly *home pages* of certain institutions) that have literally hundreds of keywords. In some cases, these long lists of key words are handed down without variation by the rest of the pages in the site. This probably has something to do with certain myths that are circulating on the way in which the search engines find and rank the pages. Some pages repeat the same keyword many times, in the hope that the search engines will place it at the top of the list.

As regards the locating of home pages, it seems that the use of query patterns to distinguish *home page* queries and treat them specifically achieves results, since experiments `USAL3` and `USAL4` showed an improvement over the others. Of these two, *Centrality* provided better results for detecting *home pages*. *Centrality* is simpler and does not discriminate backlinks, but it seems that the *home pages* are not necessarily the most prestigious.

5 Conclusions

We have described our participation in WebCLEF 2005, based on the retrieval by contents using fusion or a combination of different elements, as well as the use of coefficients from links analysis for locating *home pages*. The use of information elements such as the `TITLE` or anchor text is clearly helpful, despite the fact that the texts of many backlinks are very short. However, the `keywords` entered by the authors of the pages seem to be of little help and do not result in good results. Moreover, the coefficients based on links analysis, such as *Page-Rank* or the simple index of *Centrality*, help to locate *home pages*.

References

1. Figuerola, C.G., Zazo Rodríguez, A., Alonso Berrocal, J.L., Rodríguez, E.: Karpanta: Un motor de búsqueda para la investigación experimental en recuperación de la información. In: IBERSID 2003, Zaragoza, Spain (2003)
2. Figuerola, C.G., Zazo, Á.F., Rodríguez Vázquez de Aldana, E., Alonso Berrocal, J.L.: La recuperación de información en español y la normalización de términos. Revista Iberoamericana de Inteligencia Artificial **8**(22) (2004) 135–145
3. Beitzel, S., Jensen, E., Cathey, R., Ma, L., Grossman, D., Frieder, O., Chowdury, A., Pass, G., Vandermolen, H.: Task classification and document structure for known-item search. In: The Twelfth Text REtrieval Conference (TREC 2003), Gaithersburg, Maryland,2003. NIST Special Publication 500-255 (2003)
4. Fox, E.A., Shaw, J.A.: Combination of multiples searches. In: Overview of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-226 (1994) 243–252
5. Lee, J.H.: Combining multiple evidence from different relevance feedback methods. Technical Report, Center for Intelligent Information Retrieval (CIIR), Department of Computer Science, University of Massachusetts (1996)

6. Thompson, P.: A combination of expert opinion approach to probabilistic information retrieval, part 1: The conceptual model. Information Processing and Management **26**(3) (1990) 371–382
7. Basterr, B.T., Cottrell, G.W., Belew, R.K.: Automatic combination of multiple ranked retrieval systems. In: Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin. Ireland, 3–6 July 1994 (Special Issue of the SIGIR Forum), ACM/Springer-Verlag (1994)
8. Lee, J.H.: Analyses of multiple evidence combination. In: SIGIR '97: Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, ACM Press (1997) 267–276
9. Kraaij, W., Westerveld, T., Hiemstra, D.: The importance of prior probabilities for entry page search. In: 5th Annual International ACM SIGIR Conference, Association for Computing Machinery (2002) 27–34
10. Plachouras, V., Ounis, I., Rijsbergen, C.J.v., Cacheda, F.: University of Glasgow at the Web Track: Dynamic application of hyperlink analysis using the query scope. In: The Twelfth Text REtrieval Conference (TREC 2003), Gaithersburg, Maryland,2003. NIST Special Publication 500-255 (2003)
11. Tomlinson, S.: Robust, Web and Terabyte retrieval with Hummingbird Searchserver at TREC 2004. In: The Thirteen Text REtrieval Conference (TREC 2004), NIST Special Publication 500-261 (2004)
12. Hawking, D., Craswell, N.: Very large scale retrieval and Web search. In Voorhees, E., Harman, D., eds.: TREC: Experiment and Evaluation in Information Retrieval. MIT Press (2005) `http://es.csiro.au/pubs/trecbook_for_website.pdf` (ISBN 0262220733).
13. Yang, K., Albertson, D.: Widit in TREC 2004 genomics, hard, robust and Web tracks. In: The Thirteen Text REtrieval Conference (TREC 2004), NIST Special Publication 500-261 (2004)
14. Zaragoza, H., Craswell, N., Taylor, M., Saria, S., Robertson, S.: Microsoft Cambridge at TREC-13: Web and hard tracks. In: The Thirteen Text REtrieval Conference (TREC 2004), NIST Special Publication 500-261 (2004)
15. Farah, M., Vanderpooten, D.: Novel approaches in text information retrieval. Experiments in the Web track of TREC-2004. In: The Thirteen Text REtrieval Conference (TREC 2004), NIST Special Publication 500-261 (2004)
16. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems **30**(1–7) (1998) 107–117
17. Kleinberg, J.M., Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.S.: The Web as a graph: measurements, models, and methods. Lecture Notes in Computer Science **1627** (1999)