REINA at the WebCLEF Task: Combining evidences and Link Analysis

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

Abstract

The participation of the REINA Research Group in WebCLEF 2005 is 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, text of backlinks) and then we combine the results. For queries about *home pages*, we try to detect them with 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 coefficients.

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

Information Retrieval, Web Search, Link Analysis, Search Fusion

1 Introduction

Our participation in WebCLEF 2005 is focused in the monolingual (spanish) mixed task. This task has two goals: to find *named* web pages and *home* web pages. Every query has an only right answer: both kinds of queries are mixed, and we don't know in advance wich kind is every query.

In principle, the basic approach consists of finding the pages whose content is more similar to each query; it is hoped that the valid answer is in the first retrieves pages, and depends on the techniques applied in this search that the ranking is worse or better.

For the queries searching a *home page* we will apply some procedure that rearranges the retrieved documents list, considering, in addition to its similarity with the query, several evidences of which can be *home pages*. An additional problem is that we do not know a priori what queries or topics look for *home pages* and which not, so we will have to include some procedure that analyzes the queries and determines which persecute a *home page* and which not.

This paper is organized as follow2: section 2 describes the part of the collecion of documents which we have worked with. Section 3 decribes our aproach to task; next, we show the runs submitted anthe their results; last, conclusions are given.

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

Table 1: Blacklist for `.es` domain

2 The collection of documents

Our participation this year is limited to domain `.es` in the `EuroGov` collection. This domain has 35,168 documents; not all of these are `HTML` pages, and not always is easy to identify the format of every document. For this year, all the topics are on the `HTML` pages; the organizers provide a blacklist of 4,365 documents (in the `.es` domain) which are not `HTML`.

Nevertheless, documents in other formats nonentered in the black list exist. Thus, of 35,168 documents of the domain `.es` 8,642 does not contain the `<HTML>` tag.

Of another side, documents seems to be stripped in a size next to 64 K; in binaries files, as is the case of some PDFs, chars `chr(0)` seems to be replaced by a space (`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 some 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 diferent 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. Which of all these 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 can't find the correct page anyway.

3 Our approach

As we said before, the basic idea is to find the most similar pages to every query, and, for the *home pages* queries, rearrange the list of retrieved documents boosting those more likely *home pages*.This carry us, in addition, to analyze the queries to determine the type of these.

First part, to find the most similar pages to every query, can be solved by a classic information retrieval aproach. Nevertheless, web pages have informative elements other than the simple text which we can see at the browser's window. Thus, we can use these elements to improving the retrieval

3.1 Combining elements

The possible list of elements we can take in account in the web pages is extensive, but we focused in:

- the field `body`, which seems the most important
- the field `title`
- the contents of some `META` tags, as is the case of `Description` and `Keywords`

- the text of the backlinks, that is the links wich, in the other documents are pointing to the page tha we are analyzing.

All this elements are evidences tha we can combine to find the most similar pages to every query. There are several ways to do the fusion, or combining these elements; a first issue is to do the fusion prior or after run the query.

Our choice is to do it after; so, the procedure tha we applied is as follows:

- to build an index with every of the elements tha we take in account
- to run the query in every one of these indexes
- to combine the results achieved with every of indexes

For the first step, we have used our software *Karpanta* [5], based on the well known vector space model, and we built indexes of: body, title, meta description, meta keywords and text of backlinks. Terms weights are computed in a classic way based int $tf \times IDF$ known as *atc*. In all cases stop words (from a standard list of about 300 spanish words) were removed, and a enhanced s-stemmer was applied [6].

The size of the indexes is different, as are the fields on wich the indexes are based on. Almost all HTML pages have a field `body` (some of them only have java scripts and so on), but is not the same with the other indexes. So, 71.5 % of the pages in the `.es` domain have a field `title`, and the average size of the titles is about 40 characters; this is likely the titles are, in general, very shorts.

On the `META Description` tag, is present in only 16.9 % of the documents, with an average size of 38.6 characters. From these documents with `META Description` tag, in 7.4 % of them the content of the `META Description` tag is identical to the field `title`.

About the keywords (`META Keywords` tag), they are present in 24.7 % of the documents, with 7.7 keywords per document, in average (a keyword is not a term, but every expresion delimited with a semicolon inside the tag; so, there are keywords wich are multiword expresions).

24.7 % of the documents don't receive any link (from the documents in the collection); documents with backlinks receive an average of 9 per document. Text of these backlinks is very short (18.7 characters in average), but, perhaps, very significative.

So, it seems clear that, except the `body` field, the other elements seems to have a limited importance, as they are absents in lots of documents.

For the fusion of the list produced by every retrieval of every index, a *z-score* normalization of the similarity values [2] was performed and then the lists were merged with the *CombMNZ* algorithm [7], adapted to weight in differents ways the results obtained with every index:

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

There are several procedures of combining [7], [11],[14], [1]. Most of them are based on combining the similarity values obtaines after run the query on every of indexes; nevertheless, we can also work with the rank positions in the lists of retrieved documents in every index [12]. This algorithm has the advantage of the simplicity, as not even is necessary to normalize the similarity values.

3.2 To find *home pages*

First we must determine wich queries are about *home pages*. The concept of *home page*, nevertheless, is fuzzy; so, some of the correct answers to some queries, everybody would not consider *home pages*.

In a exploratory phase, we examined manually several *home pages* from the `.es`; specially, we examined de `title` field, as we think that a query searching for these page, probably was enough similar to the title of this one. Besides, we examined the *home page* queries used in

TREC. They are in English, but, after translated to Spanish, they can aproximate the structure and characteristics of this kind of query.

In this exploratory phase we observed some common elements in the structure of the *home page* queries. This structure lies about using certain terms in relationship with the searched *home page*. Thus, this kind of pages are entry pages to the webs of certains institutions: ministries, institutes, centers, etc. So, these terms will be present in the query [2].

Besides, they will be in certains positions inside the query, and they will go accompanied, before and later, of certain auxiliary words (articles and other connectors). This allowed us to build a set of *home page* query patterns, to which we added a simple heuristic: the presence of expresions as `home page`, `portal`, etc.

With this technique we were able to correctly identify 32 *home page* queries, 4 were erroneously considered as *home*, and 27 could not be classified.

Once identified, trough this way, the assumed *home page* queries, the results of a retrieval made with the fusion of evidences as we have seen before, were re-ranked in a way that the relevant pages most probably *home page* were in the first places.

There are several techniques to determine which retrieved pages can be *home pages*. These are not excluding techniques and they can be combined. The most known techniques are based on using two types of information: the URL page structure, and the link analysis.

Techniques based on URL structure work with the URL deep. [10] studied the statistical distribution of *home pages* in several URL deep levels, and also [2]. [13] also use techniques based on the URL length, as [15] do.

Techniques based on link analysis also are widely used. Although considered of smaller utility in the searches by content, they seem effective to retrieve *home pages* [8]. Several coefficients are used, from the simples *in* and *out-degrees* [18], to most sofistied *page-rank* [19] or *HITS* [4] algorithms.

We have tried with *Page-Rank* [3], and with *Centrality* [9], both based on backlinks.

4 Runs submitted

Our goal is to determine which elements or evidences are useful in a search based on contents; also, to test the effectiveness of coefficients based on link analysis to find *home pages*.

Official results are given in table 2. Run USAL0 acts as baseline, and it consist in queries in Spanish against the pages of the `.es` of EuroGov Collection. In this run, we work with the field `body` only.

Run USAL1 combines results of fields `body`, `title`, `META Description` and text of backlinks of every page.

Run USAL2 adds to the USAL1 the field `META Keywords`. Runs USAL3 and USAL4 try to apply specific techniques to find *home pages*. On the retrieved documents of the run USAL1, a try to detect the *home page* topics is done, and then, results are been re-ranked with Page-Rank (USAL3) and centrality (USAL4).

4.1 Evaluation

Table 2 shows the results of the official evaluation of the submitted runs. However, we have seen before somo problems about the queries (duplicateds 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 were.

4.2 Results

It seems clear that working with more elements, in addition to the `body` field, improves retrieval. This is true in the case of `title`, `META Description` and the text of the backlinks. However,

	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 2: Results of the Official Evaluation

	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 3: Unofficial Evaluation

including `META Keywords` makes worse the results. This can be surprising (some simplistics retrieval systems are based only on this field), but, if we examine the uses tha pages do of this field, we will see that, at least, it is a strange use. Table 4 shows the most used keyword expressions (not individual terms) in the `.es` domain.

Most of them are very generic expressions, little useful for searches that take place on a governmental collection. Some are included in pages also translated to English, some are directly included in English, without version in Spanish (although the language of the rest of the page is the Spanish).

A manual examination of some page of the collection shows that there are pages (specially *home pages* of certain institutions) having, literally, hundreds of keywords. In some cases, these lists of keywords are inherited with no variation by the rest of the pages of that site. Probably this has something to see with some myths that circulate on the form in which the search engines find and rank the pages. Some pages repeat a lot of times same keyword, in the hope of search engines place it in the first positions of the list.

As for the location of *home pages*, it seems that the use of patterns to distinguish *home page* queries and to treat them specifically works on, since runs USAL3 and USAL4 improves on the previous ones. Of these two, Centrality produces better results to detect *home pages*. Centrality is simpler and it does not discriminate backlinks, but it seems that the *home pages* not necessarily are the most prestigious.

5 Conclusions

We have described our participation in WebCLEF 2005, based on the retrieval by contents by means of the fusion or combination of different elements, as well as on the use of coefficients coming from the link analysis for the location of *home pages*.

The use of elements of information as the `TITLE` or the text of backlinks improves clearly the retrieval, although many pages even lack TITLE or backlinks; and although the texts of many backlinks are very short. Nevertheless, keywords introduced by the authors of the pages is from little aid and they do not produce good results.

Coefficients based on the analysis of links, like Page-Rank or the simple Centrality Coefficient, helps to locate *home pages*.

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

Table 4: Most frequent keywords in `.es`

References

[1] B. T. Basterr, G. W. Cottrell, and R. K. Belew. 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.

[2] Steve Beitzel, Eric Jensen, Rebecca Cathey, Ling Ma, David Grossman, Ophir Frieder, Abdur Chowdury, Greg Pass, and Herman Vandermolen. Task classification and document structure for known-item search. In TREC12 [16].

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[4] Mohamed Farah and Daniel Vanderpooten. Novel approaches in text information retrieval. experiments in the web track of trec-2004. In TREC13 [17].

[5] C. G. Figuerola, Á Zazo Rodríguez, J. L. Alonso Berrocal, and E. Rodríguez. Karpanta: Un motor de búsqueda para la investigación experimental en recuperación de la información. In *IBERSID 2003*, Zaragoza, Spain, 2003.

[6] 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. *Revista Iberoamericana de Inteligencia Artificial*, 8(22):135–145, 2004.

[7] E. A. Fox and J. A. Shaw. Combination of multiples searches. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 243–252. NIST Special Publication 500-226, 1994.

[8] David Hawking and Nick Craswell. Very large scale retrieval and web search. In Ellen Voorhees and Donna Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval.* MIT Press, 2005. http://es.csiro.au/pubs/trecbook_for_website.pdf

[9] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins. The web as a graph: measurements, models, and methods. *Lecture Notes in Computer Science*, 1627, 1999.

[10] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *"5th Annual International ACM SIGIR Conference*, pages 27–34. Association for Computing Machinery, 2002.

[11] Joon Ho Lee. Combining multiple evidence from different relevance feedback methods. Technical report, Center for Intelligent Information Retrieval (CIIR), Department of Computer Science, University of Massachusetts, 1996.

[12] Joon Ho Lee. Analyses of multiple evidence combination. In *SIGIR '97: Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 267–276, New York, NY, USA, 1997. ACM Press.

[13] V. Plachouras, I. Ounis, C. J. van Rijsbergen, and F. Cacheda. University of glasgow at the web track: Dynamic application of hyperlink analysis using the query scope. In TREC12 [16], page 646.

[14] P. Thompson. A combination of expert opinion approach to probabilistic information retrieval, part 1: The conceptual model. *Information Processing and management*, 26(3):371–382, 1990.

[15] Stephen Tomlinson. Robust, web anf terabyte retrieval with hummingbird searchserver at trec 2004. In TREC13 [17].

[16] *The Twelfth Text REtrieval Conference (TREC 2003), Gaithersburg, Maryland,2003.* NIST Special Publication 500-255, 2003.

[17] *The Thirteen Text REtrieval Conference (TREC 2004), Gaithersburg, Maryland (USA).* NIST Special Publication 500-261, 2004.

[18] K. Yang and D. Albertson. Widit in trec 2004 genomics, hard, robust and web tracks. In TREC13 [17].

[19] Hugo Zaragoza, Nick Craswell, Michael Taylor, Suchi Saria, and Stephen Robertson. Microsoft cambridge at trec-13: Web and hard tracks. In TREC13 [17].