SINAI at CLEF 2006 Ad Hoc Robust Multilingual Track: query expansion using the Google search engine

Fernando Martínez-Santiago, Arturo Montejo-Ráez, Miguel A. García-Cumbreras, , L. Alfonso Ureña-López
Department of Computer Science. University of Jaén, Jaén, Spain
{*dofer,amontejo,magc,laurena*}*@ujaen.es*

Abstract

This year, we have participated on Ad-Hoc Robust Multilingual track with the aim to evaluate two issues of CLIR systems. Firstly, this paper describes the method followed for query expansion in a multilingual environment by using web search results provided by the Google engine in order to increment retrieval robustness. Unfortunately, the results obtained are disappointing. The second issue reported is relative to the robustness of several usual merging algorithms. We have found that 2-step RSV merging algorithms perform better than others algorithms when geometric precision is applied.

1 Introduction

Robust retrieval has been a task in the TREC evaluation forum [5]. One of the most performant systems proposed involves query expansion through web assistance [8, 7, 6]. We have followed the approach of Kwok and his collegues and applied it for robust multilingual retrieval.

Pseudo-relevance feedback has been traditionally used to generate new queries from the results obtained from a given source query. In this way, the search is launched twice: one for obtaining first relevant documents wherefrom new query terms are extracted, and a second turn to obtain final retrieval results. This method has been found useful to resolve queries producing small result sets, and is a way to expand queries with new terms that can make the scope of the search wider. But pseudo-relevance feedback is not that useful when queries are so difficult that very few or no documents are obtained at first stage (the so-called *weak* queries). In that case, there is a straighforward solution: use another and richer collection to expand the query. Here, Internet plays a central role: it is a huge amount of web pages where almost any query, no matter how difficult it is, may be related to some subset of those pages. This approach has obtained remarkable results in monolingual IR systems evaluated in TREC conferences. Unexpectedly, in a multilingual scenario the obtained results are very poor and we think that our implementation of the approach must be tunned for CLEF queries, in spite of our conviction that an intensive tuning work is unrealistic for real-world systems. In addition, such as we suspected, the quality of the expanded terms depend on the selected language.

On the other hand, we have evaluated several merging algorithms from the perspective of robustness: round-Robin, raw scoring, normalized raw scoring, logistic regression, raw mixed 2-step RSV, mixed 2-step RSV based on logistic regression and mixed 2-step RSV based on bayesian logistic regression. We have found that round-Robin, raw scoring and methods based on logistic regression perform worse than 2-step RSV merging algorithms.

The rest of the paper has been organized into three main sections: first, we describe the experimentation framework, then we report our bilingual experiments with web-based expansion queries, and finally we describe the multilingual experiments and the way the geometric precision affects to several merging algorithms.

2 Experimentation framework

In this section we describe briefly the architecture of the multilingual system, translation approaches, query preprocessing and merging approaches.

Our Multilingual Information Retrieval System uses English as the selected topic language, and the goal is to retrieve relevant documents for all languages in the collection, listing the results in a single, ranked list. In this list there is a set of documents written in different languages retrieved as an answer to a query in a given language, English in our case. There are several approaches for this task, such as translating the whole document collection to an intermediate language or translating the question to every language found in the collection. Our approach is the latter: we translate the query for each language present in the multilingual collection. Thus, every monolingual collection must be preprocessed and indexed separately. The preprocessing and indexing tasks are described below.

2.1 Preprocessing and translation resources

In CLEF 2006 the multilingual task is made up by six languages: Dutch, English, French, German, Italian and Spanish. The pre-processing of the collections is the usual in CLIR, taking into account lexicographical and morphological idiosyncratic of every language. The pre-processing is summarized in table 1.

- English has been pre-processed as usually done in past years. Stop-words have been eliminated and we have used the Porter algorithm[11] as it is implemented in the ZPrise system.
- Dutch, German and Swedish are agglutinative languages. Thus, we have used the decompounding algorithm depicted in [10]. Stopword list and stemmer algorithm have been obtained in the Snowball site [1].
- The resources for French and Spanish have been updated by using the stop-word lists and stemmers from http://www.unine.ch/info/clef. The translation from English has been carried out by using Reverso[2] software.
- Dutch and Swedish translations have been carried out by using online FreeTrans service[3].

Table 1: Language preprocessing and translation approach

	Dutch	English	French	German	Spanish	Italian
Preprocessing	stop words removed and stemming					
Decompounding	yes	no	no	yes	no	yes
Translation approach	FreeTrans		Reverso	Reverso	Reverso	FreeTrans

Once collections have been pre-processed, they are indexed with the IR-N [14], a IR system based on passage retrieval. OKAPI model has also been used for the on-line re-indexing process required by the calculation of 2-step RSV, using the OKAPI probabilistic model (fixed empirically at $b = 0.75$ and $k1 = 1.2$) [12]. As usual, we have not used blind feedback because the improvement is very poor for these collections, the precision is even worse for some languages (English and Swedish).

[1] Snowball is a small string-handling language in which stemming algorithms can be easily represented. Its name was chosen as a tribute to SNOBOL. Available at http://www.snowball.tartarus.org

[2] Reverso is available on-line at www.reverso.net

[3] FreeTrans is available on-line at www.freetranslation.com

2.2 Merging strategies

This year we have selected the following merging algorithms: round-Robin, raw scoring, normalized raw scoring, logistic regression, raw mixed 2-step RSV, mixed 2-step RSV based on logistic regression and mixed 2-step RSV based on bayesian logistic regression:

- Round-Robin fashion. The documents are interleaved according to rank obtained for each document by means of monolingual information retrieval processing. Thus, given a multilingual collection and N languages, the first document for each monolingual retrieval list will constitute M first documents, the second document of each list will constitute the next M documents, and so on. In this case, the hypothesis is the homogeneous distribution of relevant documents across the collections. This merging process decreases precision about 40% because of the merging process [15, 18].

- Raw-scoring. This method produces a final list sorted by document score computed independently for each monolingual collection. This method works well whether each collection is searched by the same or a very similar search engine and query terms are distributed homogeneously over all the monolingual collections. Heterogenous term distribution will mean that query weights may vary widely among collections [9], and therefore this phenomenon may invalidate the raw-score merging hypothesis.

- Normalized scoring. An attempt to make document scores comparable is by normalizing in some way the document score reached for each document:

  - Given a monolingual collection, by dividing each RSV by the maximum RSV reached in such a collection:

$$RSV_i' = \frac{RSV_i}{\max(RSV)}, 1 <= i <= N \quad (1)$$

  - A variant of the previous method is to divide each RSV by the difference between the maximum and minimum document score values [17] reached for each collection:

$$RSV_i' = \frac{RSV_i - \min(RSV)}{\max(RSV) - \min(RSV)}, 1 <= i <= N \quad (2)$$

- Original 2-step RSV merging strategy consists of calculating a new RSV (Retrieval Status Value) for each document in the ranked lists at every monolingual list. The new RSV, called two-step RSV, is calculated by reindexing the retrieved documents according to a vocabulary generated from query translations, where words are aligned by meaning, i.e. each word is aligned with its translations [10]. The query is translated using an approach based on Machine Translation (MT), when available. Note that since MT translates the whole of the phrase better than word for word, the 2-step RSV merging algorithm is not directly feasible with MT. Thus, we proposed a straightforward and effective algorithm in order to align the original query and its translation at term level.

  Although the proposed algorithm to align phrases and translations at term level works well, it does not obtain fully aligned queries. In order to improve the system performance when some terms of the query are not aligned, we generate two subqueries. The first one is made up by the aligned terms only and the other one is formed with the non-aligned terms. Thus, for each query every retrieved document obtains two scores. The first score is obtained by using the 2-step RSV merging algorithm over the first subquery. In contrast, the second subquery is used in a traditional monolingual system with the respective monolingual list of documents. Therefore, we have two scores for each query, one is global for all languages and the other is local for each language. Thus we have to integrate both values. As a way to deal with partially aligned queries (i.e. queries with some terms not aligned), we have used raw mixed 2-step RSV and logistic regression:

– Raw mixed 2-step RSV method:

$$RSV_i' = \alpha \cdot RSV_i^{align} + (1-\alpha) \cdot RSV_i^{nonalign} \quad (3)$$

where RSV_i^{align} is the score calculated by means of aligned terms, as original 2-step RSV method shows.On the other hand, $RSV_i^{nonalign}$ is calculated locally. Finally, α is a constant (usually fixed to $\alpha = 0.75$).

– Logistic regression: [16] proposes a merging approach based on logistic regression. Logistic regression is a statistical methodology for predicting the probability of a binary outcome variable according to a set of independent explanatory variables. The probability of relevance to the corresponding document D_i will be estimated according to both the original score and logarithm of the ranking. Based on these estimated probabilities of relevance, the monolingual list of documents will be interleaved forming a single list:

$$Prob[D_i \ is \ rel | rank_i, rsv_i] = \frac{e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i}}{1+e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i}} \quad (4)$$

The coefficients α, β_1 and β_2 are unknown parameters of the model. The usual methods when fitting the model tend to be maximum likelihood or iteratively re-weighted least squares methods. Because this approach requires fitting the underlying model, the training set (topics and their relevance assessments) must be available for each monolingual collection. In the same way that the score and $\ln(rank)$ evidence was integrated by using logistic regression (Formula 4), we are able to integrate RSV^{align} and $RSV^{nonalign}$ values:

$$Prob[D_i \ is \ rel | \Theta] = \frac{e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i^{align}+\beta_3 \cdot rsv_i^{nonalign}}}{1+e^{\alpha+\beta_1 \cdot rsv_i^{align}+\beta_2 \cdot rsv_i^{nonalign}}} \quad (5)$$

where $\Theta = rank_i, rsv_i^{align}, rsv_i^{nonalign}$ and RSV_i^{align} and $RSV_i^{nonalign}$ are calculated as Formula 3. Again, training data must be available in order to fit the model. This is a serious drawback, but this approach allows integrating not only aligned and non-aligned scores but also the original rank of the document:

$$Prob[D_i \ is \ rel | \Theta] = \frac{e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i^{align}+\beta_3 \cdot rsv_i^{nonalign}+\beta_4 \cdot rsv_i^{local}}}{1+e^{\alpha+\beta_1 \cdot \ln(rank_i)+\beta_2 \cdot rsv_i^{align}+\beta_3 \cdot rsv_i^{nonalign}+\beta_4 \cdot rsv_i^{local}}} \quad (6)$$

where rsv_i^{local} is the local rank reached by D_i at the end of the first step, and $\Theta = rsv_i^{local}, rank_i, rsv_i^{align}, rsv_i^{nonalign}, rsv_i^{local}$.

- In addition, this year we have used bayesian logistic regression such as is implement in BBR package[4].

3 Query expansion using the Internet as resource

Expanding user queries by using web search engines such as Google has been successfully used for improving robustness of retrieval systems over collections in English language. Due to the multilinguality of the web, we have assumed that this could be extended to additional languages, though the smaller amount of non-english web pages could represent an important drawback. In figure 1 the process for query expasion by using the Internet is drawn. The process is splitted into the following steps:

1. **Web query generation**. First, we take the original query and generate a set of words that will be used to search the web. Since queries in CLIR contain title and description fields, it is important to define how terms are taken from these fields. Depending whether we consider the title field or the description field, the generation of the query varies:

[4]BBR software available at http://www.stat.rutgers.edu/ madigan/BBR.

- *From title.* Experiments expanding queries based just on the title field take all the terms in the field in lower case joined with the AND operator.
- *From description.* For those experiments where the description field is the source of terms to generate the web query, a selection of terms has to be done. For that, stop words are removed (using a different list according to the language the description is written in) and the top 5 ranked terms are taken to compose, as for the title field, an AND query. The score computed for each term to rank them obeys the following formula:

$$w_k = \frac{(F_k/D_k)^{1.5}}{\log(max\{2000, D_k\})} \tag{7}$$

where

w_k is the weight of term k

F_k is the frequency of term k (number of ocurrence in the description field)

D_k is the document frecuency of term k (number of fields the term appears in)

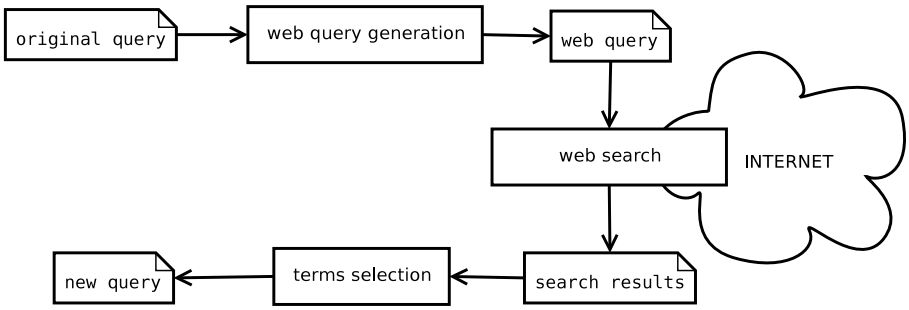

Figure 1: Using the Internet for query expansion

2. **Web search** Once the web query has been composed, the web search engine is called to retrieve relevant documents. For this, the Google API (*Application Programming Interface*) enables the use of its search engine facilities inside our programs. Thus, we can automate the process of query expansion through Google using its Java API. This web search is done specifying the language of the documents expected for the retrieval. Therefore, a filter on the language is set on the Google engine.

3. **Term selection from web results** Google returns documents in order of relevance in groups of 10 items; our implementation takes into account the 20 top ranked items (thus, the first two pages of search results). Each item points to an URL but also contains the so-called "snippet", which is a selection of text fragments from the original pages containing the terms involved in the web search (i.e. the query terms). This kind of summary is intented to let the user better follow those links that are of its real interest. In our implementation of query expansion by using web documents we have performed experiments using just the snippets as retrieved text in order to propose new query terms, and also experiments where terms are selected from full web page content (dowloading the document from the returned URL).

   In both cases (selection of terms from snippets or selection from full web pages), the final set of terms is the composite of those 60 terms with the highest frequency after discarding stop words. Of course, in the case of full web pages, the HTML tags are also conveniently eliminated. To generate the final expanded query, terms are repeated according to its frequency (normalized to that of the least frequent term in the group of 60 selected terms).

As an example of the queries generated by the described process, for a title with words "`inondation pays bas allemagne`" the resulting expansion would produce the text:

```
pays pays pays pays pays pays pays pays pays pays pays pays
pays pays pays pays pays bas bas bas bas bas bas bas bas bas bas bas
bas bas allemagne allemagne allemagne allemagne allemagne allemagne
allemagne inondations inondations inondations france france france
inondation inondation inondation sud sud cles cles belgique belgique
grandes grandes histoire middot montagne delta savent fluviales
visiteurs exportateur engag morts pend rares projet quart amont
voisins ouest suite originaires huiti royaume velopp protection
luxembourg convaincues galement taient dues domination franque xiii
tre rent commenc temp monarchie xii maritime xive proviennent date
xiiie klaas xiie ques connu or sinter ans anglophones
```

3.1 Experiments and results

For every language we have generated four different collections of queries, one without expansion and three with web-based expansion:

1. `base` – No expansion, the original query is used and its results taken as base case
2. `sd-esnp` – Expansion using the original description field for web query generation and final terms selected from snippets
3. `st-esnp` – Expansion using the original title field for web query generation and final terms selected from snippets
4. `st-efpg` – Expansion using the original title field for web query generation and final terms selected from full web pages

Results obtained are discouraging as all our expansions lead to worse measurements of both *R-precision* and *average precision*. Figures 2 and 3 show graphically values obtained when evaluating on these measures. For technical reasons the expansion of type `st-efpg` for Dutch was not generated.

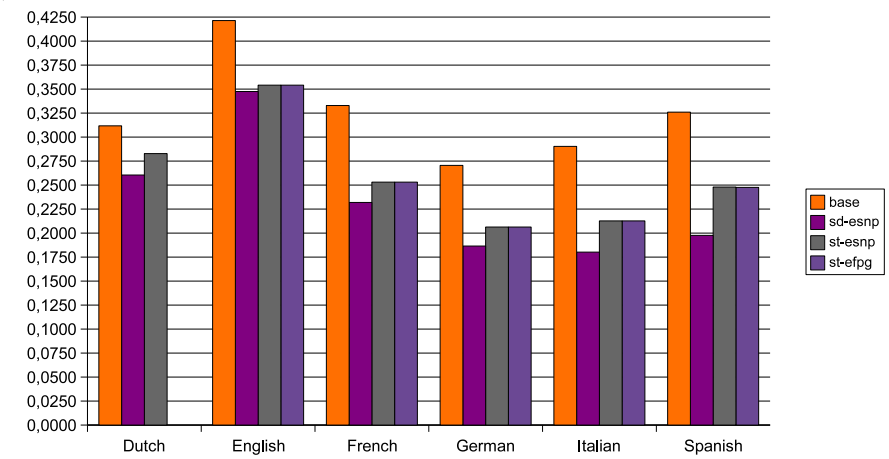

Figure 2: R-precision measurements

In a robust evaluation the key measure should the *geometric average precision* since it emphasizes the effect of improving retrieved documents on weak queries, as the task itself defines.

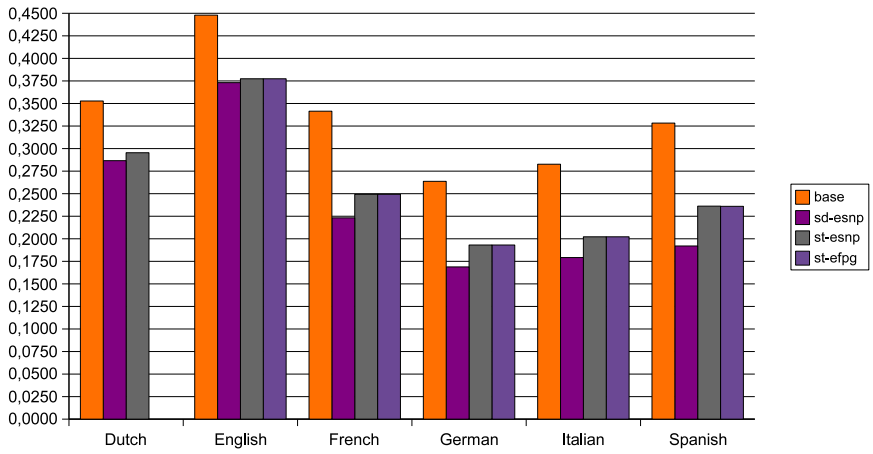

Figure 3: Average precision measurements

For future work we plan to study the value obtained on such a measure when using expanded queries and when merging retrieved items in a multilingual retrieval, as it is difficult to explain the godness of our approach on the robust task without it.

From the results above some conclusions can be extracted. The main one is that the title field is a much more suitable source of items for a web-based expansion. Indeed, for many authors the title can be considered as the set of query terms that the users should pass to a search engine. Thus, web query generation from the description field even using sophisticated formulae is, as results reflect, a worse choice when a title field is available.

The second observation is on the fact of very similar results independently on the final selection of terms, that is, it seems that the decision of taking final terms either from snippets or from full web pages text does not determine significant differences on results obtained. This issue needs further investigation since expanded queries are quite different on the last half of the selected terms (those that are less frequent) and these results make us think of the system not profiting from the full set of terms passed.

As last underlined point, we find that results depends on the language under study. We think this is due to differences on the size of existing collections of pages for each language found in the web, and that could explain the slightly better results in the case of English compared to the rest of languages.

4 Multilingual experiments

As the section 2 is depicted, the merging algorithm is the only difference between all our multilingual experiments. Table 2 show the obtained results in terms of 11-pt average precision, R-precision and the new measure geometric precision. From the point of view of the average precision, the more interesting result is the relatively poor result obtained by the methods based on machine learning. Thus, mixed 2-step RSV-LR and mixed 2-step RSV-BLR performs slightly worse than mixed 2-step RSV-LC in spite of this last approach does not use any training data. As usual, logistic regression performs better than round-Robin and raw scoring, but the difference is not as relevant as other years. Thus, we think that difficult queries are not learned as good as usual queries, probably because, given a hard query, the relation between score, ranking and relevance of a document is not clear at all, therefore machine learning approaches are not capable to learn a good enough prediction function. In the same way, this year there are not only hard queries, but also very heterogeneous queries too, from the point of view of average precision. Thus, the distribution of average precision is very smooth and it makes more difficult extracting useful information from the training data.

Table 2: Multilingual results. Raw mixed 2-step RSV is the

Merging approach	11Pt-AvgP	R-precision	Geometric Precision
round-robin	23.20	25.21	10.12
raw scoring	22.12	24.67	10.01
normalized Raw scoring	22.84	23.52	10.52
logistic regression	25.07	27.43	12.32
raw mixed 2-step RSV	27.84	32.70	15.70
mixed 2-step RSV based on LR			
mixed 2-step RSV based on BLR			

Since the 2-step RSV overcomes largely the rest of tested merging algorithms when they are evaluated by using geometric precision measure, we think that 2-step RSV merging algorithm is better suited than other merging algorithms in order to improve the robustness of CLIR systems. In this way, if we use geometric precision to evaluate de CLIR system, the difference of performance between results by using 2-step RSV and the rest of merging algorithms is higher than by using traditional 11Pt-AvP or R-precision measures.

5 Conclusions

We have reported our experimentation for Ad-Hoc Robust Multilingual track CLEF task about web-based query expansion for other languages than English. Firstly, we try to apply the expansion of queries by using web search engine such as Google. This approach has obtained remarkable results in monolingual IR systems evaluated in TREC conferences. But in a multilingual scenario the obtained results are very poor and we think that our implementation of the approach must be tunned for CLEF queries, in spite of our belief in that an intensive tuning work is unrealistic for real-world systems. In addition, such as we suspected, the quality of the expanded terms depend on the selected language. The second issue reported is relative to the robustness of several usual merging algorithms. We have found that Round-Robin, raw scoring and methods based on logistic regression performs worst from the point of view of robustness. On the other hand, 2-step RSV merging algorithms perform better than the others algorithms when geometric precision is applied. Anyway, we think that the development of a robust CLIR system does not require special merging approaches, it "only" requires good merging approaches. Maybe that other CLIR problems such as translation strategies or the development or an effective multilingual query expansion should be revisited in order to obtain such a robust CLIR model.

6 Acknowledgments

This work has been supported by Spanish Government (MCYT) with grant TIC2003-07158-C04-04.

References

[1] E. M. Voorhees: *The TREC Robust Retrieval Track*, TREC Report 2005

[2] K.L. Kwok, L. Grunfeld, P. Deng: *Improving Weak Ad-Hoc Retrieval by Web Assistance and Data Fusion*, AIRS 2005, LNCS 3689, pp. 17–30, 2005

[3] K.L. Kwok, L. Grunfeld, H.L. Sun, P. Deng: *TREC2004 Robust Track Experiments using PIRCS, 2004*, 2005

[4] L. Grunfeld, K.L. Kwok, N. Dinstl, P. Deng, 2003, *TREC2003 Robust, HARD and QA Track Experiments using PIRCS*, 2003

[5] E. M. Voorhees: *The TREC Robust Retrieval Track*, TREC Report 2005

[6] K.L. Kwok, L. Grunfeld, P. Deng: *Improving Weak Ad-Hoc Retrieval by Web Assistance and Data Fusion*, AIRS 2005, LNCS 3689, pp. 17–30, 2005

[7] K.L. Kwok, L. Grunfeld, H.L. Sun, P. Deng: *TREC2004 Robust Track Experiments using PIRCS, 2004*, 2005

[8] L. Grunfeld, K.L. Kwok, N. Dinstl, P. Deng, 2003, *TREC2003 Robust, HARD and QA Track Experiments using PIRCS*, 2003

[9] S.T. Dumais. Latent Semantic Indexing (LSI) and TREC-2. *Proceedings of TREC'2, volume 500-215, pages 105-115, Gaithersburg*, 1994. NIST, D. K. Harman.

[10] F. Martinez-Santiago, L.A. Ureña, and M. Martin. A merging strategy proposal: two step retrieval status value method. *Information Retrieval, vol. 9, issue 1, 71-93*, Jan 2006.

[11] M.F. Porter. An algorithm for sufix stripping. *Program 14, pages 130-137*, 1980.

[12] S. E Robertson, S. Walker., and M. Beaulieu. Experimentation as a way of life: Okapi at TREC. *Information Processing and Management, vol. 1, 95-108*, 2000.

[13] J. Savoy. Cross-Language information retrieval: experiments based on CLEF 2000 corpora. *Information Processing and Management, vol 39, 75-115*, 2003.

[14] F. Llopis, H. Garcia Puigcerver, Mariano Cano, Antonio Toral, Hector Espi. *IR-n System, a Passage Retrieval Architecture. TSD, 57-64*, 2004.

[15] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. *Proceedings of the 18th International Conference of the ACM SIGIR'95, pages 21-28, New York*, 1995. The ACM Press.

[16] A. Calve and J. Savoy. Database merging strategy based on logistic regression. *Information Processing and Management, 36:341-359*, 2000.

[17] A. L. Powell, J. C. French, J. Callan, M. Connell, and C. L. Viles. The impact of database selec- tion on distributed searching. *The ACM Press., editor, Proceedings of the 23rd International Conference of the ACM-SIGIR'2000, pages 232-239, New York.* 2000.

[18] E. Voorhees, N. K. Gupta, and B. Johnson-Laird. The collection fusion problem. *D. K. Harman, editor, Proceedings of the 3th Text Retrieval Conference TREC-3, volume 500-225, pages 95-104, Gaithersburg*, 1995. National Institute of Standards and Technology, Special Publication.