MIRACLE's Approach to Multilingual Web Retrieval

Ángel Martínez-González[1,3], José Luis Martínez-Fernández[2,3]
César de Pablo-Sánchez[2], Julio Villena-Román[2, 3]
Luis Jiménez-Cuadrado[2], Paloma Martínez[2], José Carlos González-Cristóbal[1, 3]

[1] Universidad Politécnica de Madrid
[2] Universidad Carlos III de Madrid
[3] DAEDALUS - Data, Decisions and Language, S.A.

amartinez@daedalus.es, jmartinez@daedalus.es,
cdepablo@inf.uc3m.es, jvillena@daedalus.es
luis.jimenez@uc3m.es, paloma.martinez@uc3m.es,
jgonzalez@dit.upm.es

Abstract

For MIRACLE participation on WebClef 2005, a set of independent indexes was constructed for each top level domain of the EuroGOV collection. Each of these indexes contains information extracted from the document, like URL, title, keywords, detected named entities or HTML headers. These indexes are queried to obtain partial document rankings, which are combined with various relative weights to test the value of each index.

The trie based indexing and retrieval engine developed by the MIRACLE team is now fully functional and has been adapted to the WebClef environment and employed in this campaign. Other tools, such as the Named Entities Recognizer based on a finite automaton, have also been developed.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.2 Information Storage; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software. E.1 [Data Structures]. E.2 [Data Storage Representations]. H.2 [Database Management].

Keywords

Linguistic Engineering, Information Retrieval, Trie Indexing, World Wide Web, known- item search, Named Entity.

1 Introduction

Linguistic heterogeneity makes the Web a very appropriate setting to evaluate cross-language Information Retrieval systems. Furthermore, web search engines face some challenges not found in other Information Retrieval tasks. The most obvious one is the great volume of collections, but other should also be cited such as heterogeneity of formats, variation in quality of documents, redundancy of documents and the need to consider web structure (hyperlinks) and metadata.

WebCLEF 2005 focuses on known-item search, i.e. retrieving a concrete page already known to exist in the collection. For this purpose, the collection used is EuroGOV, consisting of documents from European governmental sites of up to 17 top level domains. A set of topics has been collaboratively built by the participants.

The MIRACLE team is made up of three university research groups located in Madrid (UPM, UC3M and UAM) along with DAEDALUS, a company founded in 1998 as a spin-off of two of these groups. DAEDALUS is a leading company in linguistic technologies in Spain and is the coordinator of the MIRACLE team. This is the third participation in CLEF, after years 2003 and 2004 [2],[6],[9],[15],[16]. As well as bilingual, monolingual and cross lingual tasks, the team has participated in the ImageCLEF, Q&A, WebCLEF and GeoCLEF tracks.

The MIRACLE team objectives for this first participation in WebCLEF were to adapt and test our existing tools to a web environment, so as to procure a flexible set of instruments to extract and mix the information present in a web document (see Section 3 for more detail on the developed tools). Secondly, to evaluate the relative relevance of several of this information sources such as document URL, title, keywords, detected named entities or HTML headers. The MIRACLE group has taken part in the two main tasks (Mixed Monolingual and Multilingual).

2 Experiment design

For MIRACLE participation on WebClef 2005 we decided not to follow a full text approach. Instead, a set of independent indexes was constructed for each top level domain of the collection. Partial results are obtained by applying the probabilistic ranking formula BM25 [12] to these indexes. Finally, these partial results are combined to get the final result. In different experiments, different weights are given to each set of partial results, so as to evaluate the relative importance of the different information sources that have been indexed.

The generated indexes were the following:

- H1 index, containing document titles and H1 HTML headers (if you are not familiar with the HTML standard, see [11]).
- H2 index, containing headers H2 to H6.
- PN index, containing named entities (proper nouns) found by a detection module.
- Ky index, document keywords given in a META HTML element.
- Url, containing parsed parts of the document url, removing the querystring and taking characters such as '.' ,'/' or '–' as delimiters.

Consequently, the total number of indexes was 85 (5 indexes/domain * 17 domains). Another index called LINKS was initially planned but finally not included due to lack of time. This index contained the words in the anchor (<A>) elements of documents that point to the indexed document. Note that, unlike the other indexes, LINKS needs two passes over the collection, so that links pointing to a document not in the collection are discarded. Although this index was not included, the tool for building it is available for future participations.

3 Developed tools

MIRACLE toolbox [3] consists of a set of independent modules that perform extraction (XML parsing), preprocessing (word segmentation, filtering, stop word removing, stemming), indexing and retrieval of documents. In last year participation, a trie based indexing and retrieval engine was under development, but not yet finished, so a Xapian based front-end was used. This trie base engine is now fully functional and has been used for all WebCLEF experiments (an also in MIRACLE participation in other tracks). Some of the functional characteristics of this engine are:

- Several variants of probabilistic and vectorial ranking formulas can be selected, as an option of the retrieval program, with no need of reindexing. There is only one index format, which contains all the necessary information for each ranking algorithm.
- Indexing time, which is the most critical factor limiting the number of different experiments that can be performed in the available time, is optimized rather than retrieval times or index size. An important improvement has been achieved in this aspect over the previously used Xapian based tool.
- The index can be incrementally built. Deletion of terms or documents is in principle also supported, but inefficient.
- Relevance feedback is also supported, although not used in these experiments.

The other tools created by MIRACLE for WebCLEF are explained bellow, in the order they are applied to the collection:

- Document extraction: the collection is given in a few huge files with a format close to XML. The documents must be extracted.
- HTML parser: based on the El-Kabong HTML processing library [2]. The content or attributes of special tags such as headers, anchors or META tags are extracted. The body is then extracted to plain text format.

- Named Entities Recognition: named entities are filtered from plain text using a multilingual recognizer in current development. Recognition is based on the evaluation of predicates in a Finite State Automaton. We have explicitly considered Spanish, Portuguese, Italian, French, English, Swedish and Dutch. Simple and multiword proper nouns are detected by means of cues such as capitalization, words that introduce named entities ("Sr.", "president", "river"), connectors ("van", "de") and punctuation signs ("Paris-Dakar", "Madigan's", etc.).

  After WebCLEF we have evaluated the tool with data from the CONLL 2002 shared task and achieved the following results, only for recognition:

Table 1: Evaluation of the Named Entities Recognition module

	Precision	Recall
Spanish	80.49%	88.70%
Dutch	66.25%	60.38%

- Indexing and ranking: the above explained engine is used. BM25 formula is used.
- Combination of partial results: relevance rankings from different indexes are mixed by means of an ad-hoc script that calculates the average relevance allowing to easily assign different weights to different indexes.
- Query language detection: in the case of the baseline mixed monolingual run, no metadata such as the target language of the query was allowed to be used, so this module tries to guess the target language from the words of the query title. For the multilingual task we have used a list of stop words, while for the mixed monolingual task a list of locations and names of the inhabitants of a country or region. A simple vote algorithm has been used.

4 Description of the submitted runs

The MIRACLE team has taken part in the two main tasks (Mixed Monolingual and Multilingual), submitting five runs for each one of them. A baseline run, using no metadata is mandatory. The other four runs (which will be referred as extended in this paper) use supplied metadata (the target domain). In the baseline runs, the language identification tool is employed to guess the target language from the words in the query title. For each query, only the indexes of the top level domain corresponding to the target language and the international INT domain are queried. For example, if the target language of a query is known of has been guessed to be Spanish, only ES and INT domains are considered, even though there are documents in Spanish in the other domains.

In the five Monolingual runs submitted, partial results were combined in the following ways:

- Monobase: this is the baseline run. Relevance of documents is averaged over the five partial results, giving al of them the same weight.
- MonoExt: extended run, combining the results in the same way as in MonoBase.
- MonoExtH1PN: extended run; only H1 and PN indexes are considered, giving both of them the same weight.
- MonoExtUrlKy: extended run; only Url and Ky indexes are considered, giving both of them the same weight.
- MonoExtAH1PN: extended run. All indexes are considered for retrieval, but the H1, PN and Ky indexes are considered more relevant than the rest, so a weight factor with value 2 is applied for these partial lists.

In the Multilingual runs Multibase, MultiExt, MultiExtH1PN, MultieExtUrlKy and MultiExtAH1PN, partial results are mixed in the same way as in the corresponding monolingual runs. All this information is summarized in the table below.

Table 2: Weight distribution for the different experiments

			Weight of partial results				
Run name	**Mono/Multilingual**	**Metadata**	**H1**	**H2**	**PN**	**Ky**	**Url**
Monobase	Monolingual	None	1	1	1	1	1
MonoExt	Monolingual	Target domain	1	1	1	1	1
MonoExtH1PN	Monolingual	Target domain	1	0	1	0	0
MonoExtUrlKy	Monolingual	Target domain	0	0	0	1	1
MonoExtAH1PN	Monolingual	Target domain	2	1	2	2	1
Multibase	Multilingual	None	1	1	1	1	1
MultiExt	Multilingual	Target domain	1	1	1	1	1
MultiExtH1PN	Multilingual	Target domain	1	0	1	0	0
MultiExtUrlKy	Multilingual	Target domain	0	0	0	1	1
MultiExtAH1PN	Multilingual	Target domain	2	1	2	2	1

5 Evaluation of results

The following graphics are based on the evaluation results provided by WebCLEF organizers. The different parameters included in these results will be explained bellow as they appear in a figure.

The average success at n is defined as the portion of topics where the known-entity was found at a rank less than or equal to n. The following figures show this average success rate as a function of n for the monolingual and multilingual tasks.

Figure 1: Average success in monolingual runs

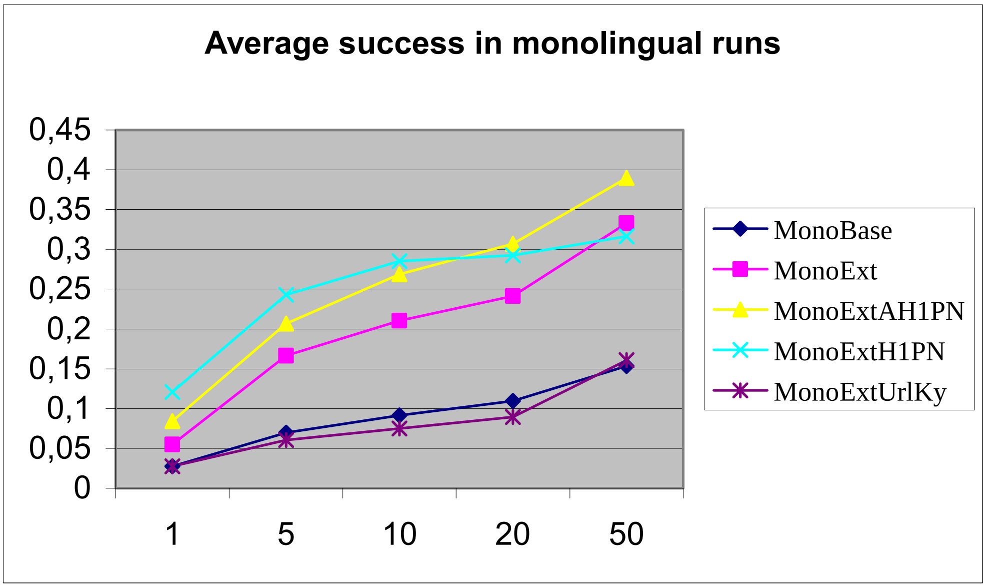
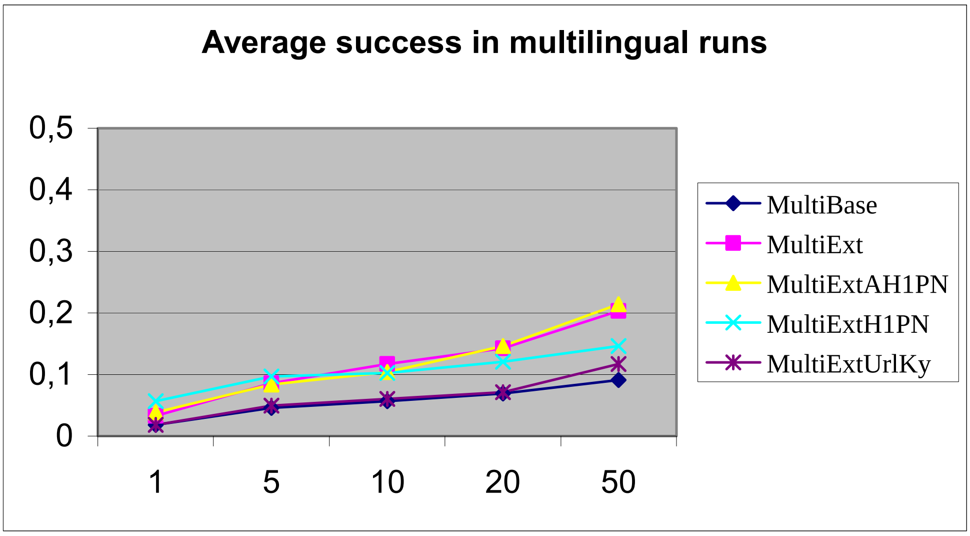


Figure 2: Average success in multilingual runs



The expected conclusion is confirmed: titles and named entities are the most valuable sources of information to find known-items. In Multilingual runs results are worse and the effect of different combinations of results are not so significant.

The Mean Reciprocal Rank (MRR) is 1 divided by the rank given to the known-entity or 0 if the relevant document has not been retrieved. The parameter called DFA is defined as the difference between the MRR score for a given topic and the average MRR score over the submitted runs of all participants. In the figures bellow, DFA values as a function of the topic are given for the best extended monolingual and multilingual runs: MonoExtH1PN and MultiExtAH1PN.

Figure 3: DFA values for MonoExtH1PN

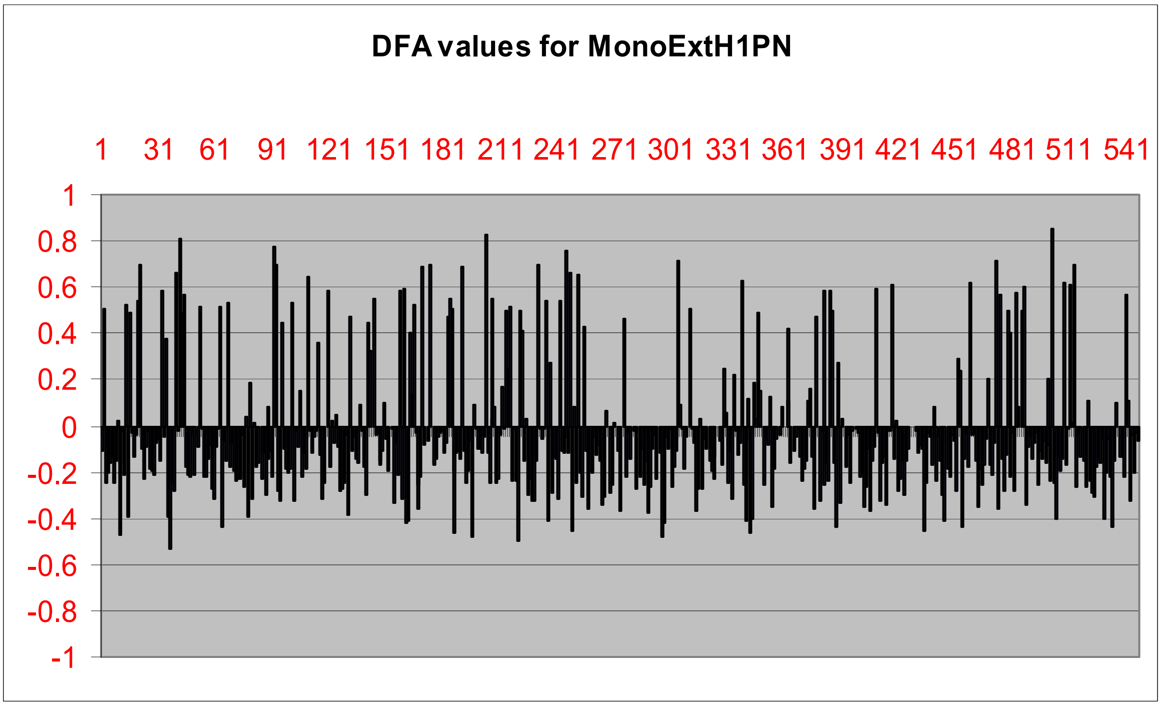
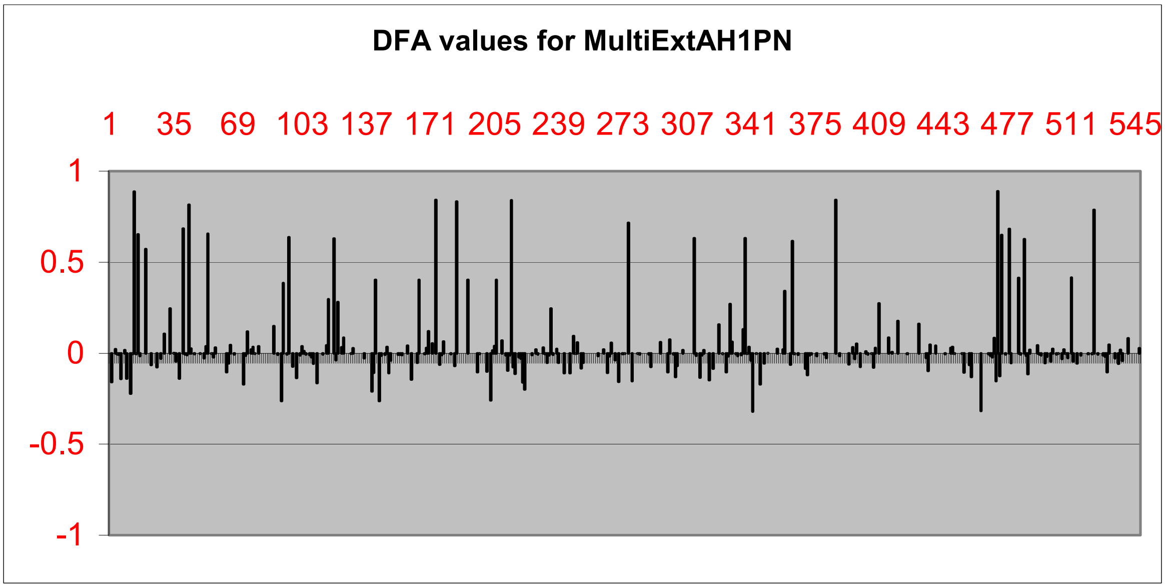


Figure 4: DFA values for MultiExtAH1PN



The DFA value averaged over all topics is -0,04 for MonoExtH1PN and +0,02 for MultiExtAH1PN: both of them are very close to 0, the mean value over runs submitted by all participants. Our results in the multilingual task are, although worse in absolute terms than the monolingual results, better if considered relative to the other participants, even though our approach was quite simple, with no query or document translation. Elements without translation such as named entities are less noisy and especially valuable for known-item search.

Although not shown in the figures above, our results were rather variable with the target language of the topic. The results in languages such as Greek or Russian were much poorer than other languages, even though the techniques used are language independent (with the partial exception of named entities recognition). This suggests we have had some sort of problem with character sets and encodings, which should be corrected for future participations.

6 Conclusions and future work

Obviously, in this first year of WebCLEF track, there were no previous results available and the selection of experiments was somehow blind. Nevertheless the foundations for future campaigns have been settled and several valuable conclusions have been drawn. We have at our disposal a set of software tools that we plan to use and further improve in future campaign in order to pursue more ambitious aims.

We believe that a full text index, combined appropriately with the more specific indexes would probably improve the results. In the next campaign, we are also planning to introduce some sort of query translation mechanism. Another improvement would be to consider the hyperlink structure of the collection; a voting algorithm could be used to estimate the relative importance of web pages and this way detect home pages. Finally, we are considering experimenting with automatic web classification using neural networks.

Acknowledgements

This work has been partially supported by the Spanish R+D National Plan, by means of the project RIMMEL (Multilingual and Multimedia Information Retrieval, and its Evaluation), TIN2004-07588-C03-01.

Special mention to our colleagues of the MIRACLE team should be done (in alphabetical order): Ana María García-Serrano, José Carlos González-Cristóbal, Ana González Ledesma, José Miguel Goñi-Menoyo, José Mª Guirao, Sara Lana-Serrano and Antonio Moreno Sandoval.

References

[1] Aoe, Jun-Ichi; Morimoto, Katsushi; Sato, Takashi. An Efficient Implementation of Trie Structures. Software Practice and Experience 22(9): 695-721, 1992.

[2] El-Kabong HTML. A speedy HTML processing library. On line http://www.ekhtml.sourceforge.net/ [visited 28/07/2005].

[3] Goñi-Menoyo, José M; González, José C.; Martínez-Fernández, José L.; and Villena, J. MIRACLE's Hybrid Approach to Bilingual and Monolingual Information Retrieval. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491, pp. 188-199. Springer, 2005 (to appear).

[4] Goñi-Menoyo, José M.; González, José C.; Martínez-Fernández, José L.; Villena-Román, Julio; García-Serrano, Ana; Martínez-Fernández, Paloma; de Pablo-Sánchez, César; and Alonso-Sánchez, Javier. MIRACLE's hybrid approach to bilingual and monolingual Information Retrieval. Working Notes for the CLEF 2004 Workshop (Carol Peters and Francesca Borri, Eds.), pp. 141-150. Bath, United Kingdom, 2004.

[5] Goñi-Menoyo, José Miguel; González-Cristóbal, José Carlos and Fombella-Mourelle, Jorge. An optimised trie index for natural language processing lexicons. MIRACLE Technical Report. Universidad Politécnica de Madrid, 2004.

[6] Martínez-Fernández, José L.; García-Serrano, Ana; Villena, J. and Méndez-Sáez, V.; MIRACLE approach to ImageCLEF 2004: merging textual and content-based Image Retrieval. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491. Springer, 2005 (to appear).

[7] Martínez, José L.; Villena, Julio; Fombella, Jorge; G. Serrano, Ana; Martínez, Paloma; Goñi, José M.; and González, José C. MIRACLE Approaches to Multilingual Information Retrieval: A Baseline for Future Research. Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 210-219. Springer, 2004.

[8] Martínez, J.L.; Villena-Román, J.; Fombella, J.; García-Serrano, A.; Ruiz, A.; Martínez, P.; Goñi, J.M.; and González, J.C. (Carol Peters, Ed.): Evaluation of MIRACLE approach results for CLEF 2003. Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway.

[9] de Pablo, C.; Martínez-Fernández, J. L.; Martínez, P.; Villena, J.; García-Serrano, A. M.; Goñi, J. M.; and González, J. C. *miraQA*: Initial experiments in Question Answering. CLEF 2004 proceedings (Peters, C. et al., Eds.). Lecture Notes in Computer Science, vol. 3491. Springer, 2005 (to appear).

[10] Porter, Martin. Snowball stemmers and resources page. On line http://www.snowball.tartarus.org. [Visited 13/07/2005]

[11] Ragget, D.; Le Hors A. and Jacobs I. (Ed.). HTML 4.01 Specifications. W3C Recommendation 24 December 1999. On line http://www.w3.org/TR/html4/ [visited 15/07/2005].

[12] Robertson, S.; Walker, S.; Hancock-Beaulieu, M.M. and Gatford, M. Okapi at trec 3. Text Retrieval Conference, 2003.

[13] SYSTRAN 5.0 translation resources. On line http://www.systransoft.com. [Visited 13/07/2005]

[14] University of Neuchatel. page of resources for CLEF (Stopwords, transliteration, stemmers, …). On line http://www.unine.ch/info/clef/. [Visited 13/07/2005]

[15] Villena, Julio; Martínez, José L.; Fombella, Jorge; G. Serrano, Ana; Ruiz, Alberto; Martínez, Paloma; Goñi, José M.; and González, José C. Image Retrieval: The MIRACLE Approach. Comparative Evaluation of Multilingual Information Access Systems (Peters, C; Gonzalo, J.; Brascher, M.; and Kluck, M., Eds.). Lecture Notes in Computer Science, vol. 3237, pp. 621-630. Springer, 2004.

[16] Villena-Román, J.; Martínez, J.L.; Fombella, J.; García-Serrano, A.; Ruiz, A.; Martínez, P.; Goñi, J.M.; and González, J.C. (Carol Peters, Ed.); MIRACLE results for ImageCLEF 2003. Working Notes for the CLEF 2003 Workshop, 21-22 August, Trondheim, Norway.

[17] Xapian: an Open Source Probabilistic Information Retrieval library. On line http://www.xapian.org. [Visited 13/07/2005]