LIC2M experiments at ImageCLEF 2004

Romaric Besançon, Patrick Hède, Pierre-Alain Moellic, Christian Fluhr
CEA-LIST/LIC2M
{*romaric.besancon,patrick.hede,pierre-alain.moellic,christian.fluhr*}*@cea.fr*

Abstract

For its first participation in the ImageCLEF campaign, the LIC2M participated in both the ad hoc task and the medical task. Using our cross-language information retrieval system and content-based image retrieval system, our goal was to perform some first experiments on merging the results of the two systems. The results show that the performance of each system highly depends on the corpus and the task: with the systems we used, text retrieval alone performs better on the ad hoc task, whereas image retrieval alone performs better on the medical task.

1 Introduction

The purpose of the ImageCLEF campaign is to study cross-language image retrieval, which includes the study of cross-language text retrieval on corpora that are specially associated with images (captions for instance) and the interactions between text and image retrieval. Since the LIC2M develops both cross-language text retrieval systems and content-based image retrieval systems, our goal was to perform some first experiments on merging strategies to integrate the information retrieved from both systems.

The goal of the ImageCLEF ad hoc task is to retrieve relevant images based on a text query. In this case, we tried to improve the results by merging the results of the cross-language text retrieval with a content-based image retrieval, using the example image given with the topic.

The goal of the ImageCLEF medical task is to retrieve relevant images based on an image query. The medical CasImage corpus is composed of a series of medical cases, associated with a text description and a set of images such as x-rays or scans. The task is to retrieve images similar to the query image with respect to its modality and the anatomic region shown. We tried, for this task, to improve the results by using an automatic feedback on the text description associated with images retrieved by a CBIR system.

We present in section 2 the retrieval systems for text and image. We then present the strategies used for the ad hoc task and the medical task and their results in section 3 and 4 respectively.

2 Retrieval Systems

2.1 Cross-language Text Retrieval System

The cross-language text retrieval system used for these experiments is the same as the one used for the CLEF multilingual task, and a more detailed description can be found in the working notes corresponding to this task or in the proceedings of the CLEF 2003 campaign [1]. The system has not been specially adapted to work on the text of the ImageCLEF corpora, and has simply been used as is. This system is a weighted boolean search engine based on a linguistic analysis of the query and the documents. Its basic principle is briefly described here.

Document processing The documents are processed through a linguistic analyzer, that performs in particular a part-of-speech tagging, a lemmatization, and extracts compounds and named entities from the text. All these elements are indexed into inverted files. For both the StAndrews and CasImage corpora, no special treatment has been performed to take into account the structure of the documents (such as photographer's name, location, date for the captions and description, diagnosis, clinical presentation in the medical cases): all fields have been taken as a single text to be analyzed.

Query processing The query is first processed through a similar analyzer (corresponding to the query language) to extract the informative elements of the text. These elements are used as query "*concepts*". Each concept is reformulated into a set of *search terms*, either using a monolingual expansion dictionary (that introduces synonyms and related words), or using a bilingual dictionary, depending on the index languages.

Search and merging Each search term is searched in the index, and documents containing the term are retrieved. All retrieved documents are then associated to a *concept profile*, indicating the presence of query concepts in the document. Documents sharing the same concept profile are clustered together, and a weight is associated to each cluster according to its concept profile and to the weight of the concepts (the weight of a concept depends on the weight of each of its reformulated term in the retrieved documents). The clusters are sorted according to their weights and the first 1000 documents in this sorted list are retrieved.

2.2 Content-base Image Retrieval System

For Image retrieval, we used a system developed at our lab, the LIC2M, called PIRIA (Program for the Indexing and Research of Images by Affinity)[2]. A user query is submitted to the system, which returns a list of images ranked by their similarity to the query image. The similarity is obtained by a metric distance that operates on every image signature. These indexed images are compared according to several classifiers : principally *Color*, *Texture* and *Form* if the segmentation of the images is relevant. The system takes into account geometric transformations and variations like rotation, symmetry, mirroring, etc. PIRIA is a global one-pass system, feedback or "relevant/non relevant" learning methods are not used.

Color Indexing PIRIA uses a global normalized color histogram. The choice of the color space is very important for a good color division. The model based on *Hue, Saturation and Value* is used to obtain a strong semantic content. Global histogram is used for the global image or after the segmentation of the image in several blocks. Splitting the image by blocks enables computation of spatial relationship. A more complex color analysis can be used with a region based segmentation. Color information of each region are mixed with form analysis (Fourier descriptors). The distance uses for the color indexing is a classical L1 norm.

Texture Indexing A global texture histogram is used for the texture analysis. The histogram is computed from the Local Edge Pattern descriptors [3]. These descriptors describe the local structure according to the edge image computed with a Sobel filtering.

Merge of results The merging of results from several indexers is computed with a boundary fusion based on the position of the result's images.

3 Ad hoc task

For the ad hoc task, we used topics in English and French. For each of these topic languages, we submitted two runs. The first one (lic2mSA*1t) was a simple text retrieval, with no use of the

image retrieval system. The second one (lic2mSA2*ti) uses a simple merging strategy integrating the results of both text and image retrieval: in this case, the image used for the image retrieval was the example image provided for each topic. The merging strategy was quite straightforward: each image is given a score that is a weighted sum of the scores given by each retrieval systems.

The results of the four runs[1] are presented in Figure 1 (precision/recall graph) and Table 1.

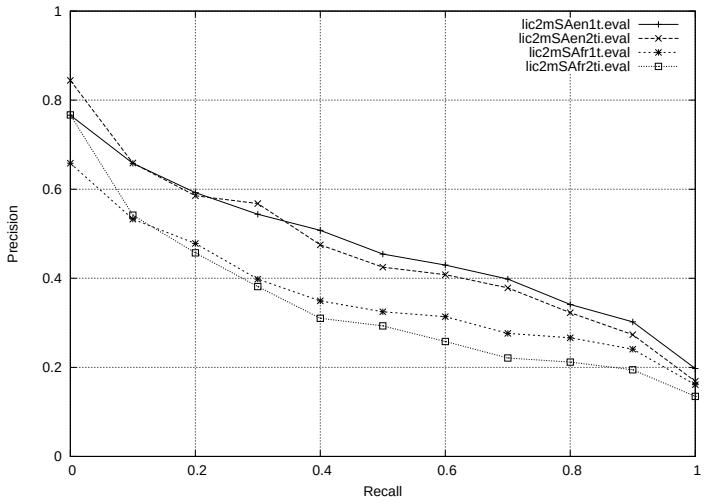

Figure 1: Precision/Recall results for the ad hoc task

	lic2mSAen1t	lic2mSAen2ti	lic2mSAfr1t	lic2mSAfr2ti
avg_p	0.42	0.423	0.314	0.295
relret	757 (91.3%)	674 (81.3%)	720 (86.8%)	565 (68.1%)

Table 1: Results for the ad hoc task: average precision, number of relevant document retrieved (with percentage)

From these results, the merging strategy using both text and image does not show much better results than the direct text search. In the case of English topics, a little improvement of the average precision is noticed (not significant), which is mostly the effect of a reordering of retrieved documents (the total number of relevant documents retrieved actually decreases). This is mainly due to the fact that the image retrieval does not perform well on this corpus (indeed, the images need a complex local analysis - based on interest points) The image retrieval alone (using the example images from the queries) retrieves only 122 relevant images out of the 829 relevant images of the `partial-isec-total` assessments. Of these 122 images, only 8 were not found by the original text retrieval, based on English topics (11 for French topics). Hence, the merging strategy seems to give too much importance to the image result and add noise to the text retrieval (removing relevant images retrieved). Nevertheless, this merging causes a reordering of relevant documents already retrieved that seems to be interesting (at least in the case of English topics).

Further experiments for merging the results of text and image are planned to try to minimize the introduction of non-relevant images in the retrieval results.

The image indexers will also be adapted to treat images such as the old photographs of the StAndrews collection: this image base is particularly difficult for the kind of image indexers we used in this experiments since most of the images are old photographs that are in a kind of

[1]The official results for the merging strategy (lic2mSA*2ti) are erroneous due to a misordering of the queries in the submitted runs (only 17 queries were taken into account out of the 25).

monochrome color (with not always the same tone), so that a color segmentation of the image cannot be performed to identify the interesting elements of the images.

4 Medical task

For the medical task, we submitted two runs. The first one (lic2mCA1i) was a simple image retrieval search from the image query. The second one (lic2mCA2it) is based on the first image retrieval search and implements an automated feedback using text information contained in the cases associated with the retrieved images.

4.1 Text feedback strategy for the CasImage corpus

The process of this feedback is the following (a schema presenting the outline of this feedback strategy is proposed in Figure 2):

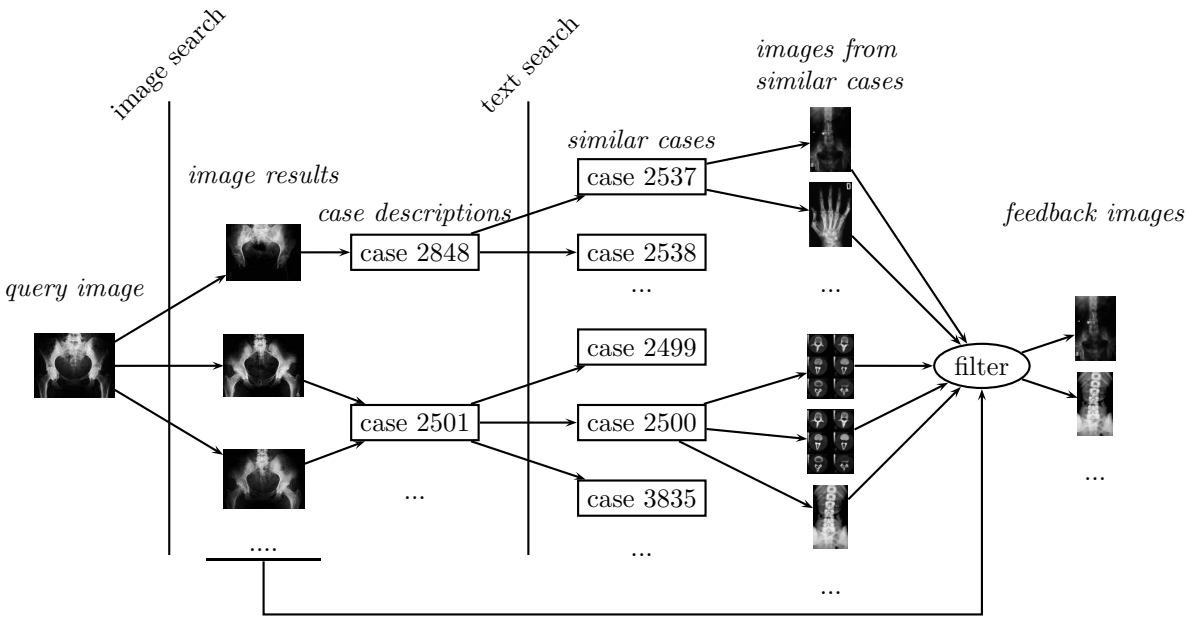

Figure 2: Text feedback strategy for the medical task

1. we first take the images retrieved by the CBIR system: these images are given a score by the CBIR system (we call it the *image score*, denoted s_i);

2. we collect the cases associated with the images retrieved by the CBIR system (we used only the top five cases for this feedback);

3. we then use these cases as queries to the text retrieval system to retrieve similar cases, based on the textual description of the cases: we retrieve the 20 most similar cases. These cases are given a score by the text retrieval system (*text score*, denoted s_t);

4. we collect the images associated to the cases retrieved by the text retrieval system: these images are candidate images for feedback;

5. since the images retrieved must have the same modality than the query image, we filter these images associated to the similar cases by their similarity to the corresponding images

collected in step 1[2]. The similarity with the original image give a score to the candidate images (*filtering score*, denoted s_f);

6. The set of feedback images is then used to enrich the first set of retrieved images (step 1), either by increasing the score of an already retrieved image (function of its image score, the text score and the filtering score) or by adding new images, with an associated score that is a function of the image score of the image that lead to the new image, the text score and the filtering score of the new image. Since the scoring of the different systems are not easily comparable, the merging of the three scores is not obvious: we used in the submitted runs an arbitrary function defined as follows: if the image was already retrieved the score is $\alpha \times s_i + (1 - \alpha) \times f(s_t, s_f)$, otherwise, the score attributed is $\alpha \times g(s_i, s_t, s_f)$, where $f(s_t, s_f)$ and $g(s_i, s_t, s_f)$ are weighted sums of the different scores. The α parameter has been introduced to make sure that images retrieved by the first step are still given more importance (in the experiments, $\alpha = 0.9$).

4.2 results

The results of both runs are presented in Figure 3 (precision/recall graph) and Table 2.

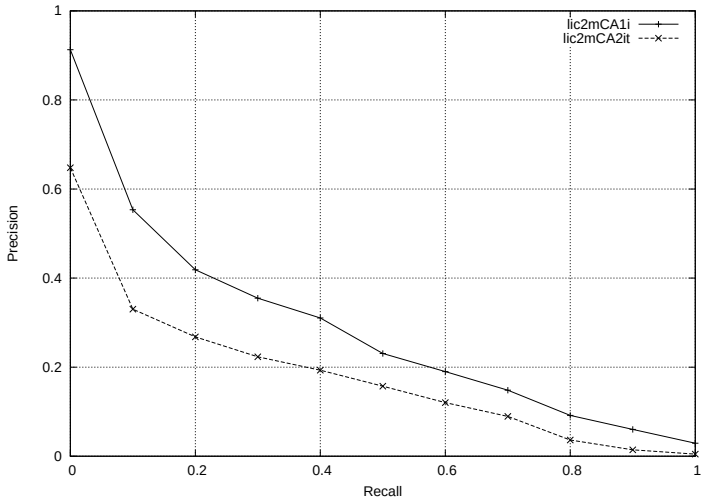

Figure 3: Results for the medical task

	lic2mCA1i	lic2mCA2it
avg_p	0.278	0.158
R_p	0.303	0.205
relret	2135 (70.3%)	2127 (70.0%)

Table 2: Results for the ad hoc task: average precision, R-precision, number of relevant document retrieved (with percentage)

In this case, the image retrieval alone performs better than the use of the text retrieval for feedback and enrichment of the retrieved images.

A deeper analysis of the feedback process show that this process produces 17656 image candidates for feedback (step 4: images from similar cases), in which 3195 images were already found

[2]We could have used directly the query image for filtering: the impact of such a choice should be studied.

by direct image retrieval (of which 829 were relevant images) and in the 14461 other images, only 243 are relevant images.

Furthermore, the merging strategy does not succeed in including these 243 documents: only 74 new relevant documents are added, and on the other hand, 82 relevant documents previously included in the initial retrieval are lost. The scoring function we tested introduce too many non-relevant images. Further testing will be performed on the scoring function, and on the number of cases to consider for feedback and the number of similar cases to look at.

Also, the task imposes that the retrieved images are of the same modality than the query images. Hence, a general similarity on the textual description of the cases is not enough: it can retrieve cases relative to the same kind of pathology but it is not obvious that the images associated to these cases will be similar to the original image We tried to avoid this problem using a second step of image similarity, but a deeper analysis of the text could be needed so that informations on the image modality and anatomic region are extracted from the case description.

Another possible reason for these results is that our text retrieval system is very general. A specialized corpus such as this medical corpus contains many technical words that are treated by the system as unknown words. A more adapted processing of the medical text, giving special importance to terms such as disease names, anatomic regions, medical acts should increase the relevance of the cases similarity.

5 Conclusion

These first experiments in the ImageCLEF campaign are very interesting: with the same two general purpose systems (no particular adaptation of the systems was made for the two tasks), the results lead to very different conclusions according to the task and corpus.

The ad hoc task with the StAndrews collection of old photographs is not well adapted the kind of image indexers we used, that relies mostly on color for segmentation. On the other hand, this task is easier for text retrieval, since the descriptions of the images in the captions are small and precise and the elements in the queries are often found as is in the documents (even without treating the structure of the captions).

The medical task offers a better field for image retrieval, the images being "easier" to index (at least, to separate the images by their modality, quite different in nature and colors) but in that case, and given the particularity of the task and the specialization of the corpus, the feedback strategy using text information did not improve the results. More experiments should be undertaken to improve the feedback strategy. Other experiments using a training process to associate text to the medical images and use the results for adding text to the image query (textual query expansion) can also be imagined.

References

[1] Romaric Besançon, Gaël de Chalendar, Olivier Ferret, Christian Fluhr, Olivier Mesnard, and Hubert Naets. The LIC2Mś CLEF 2003 system. In *Working Notes for the CLEF 2003 Workshop*, Trondheim, Norway, 21-22 August 2003.

[2] Magali Joint, Pierre-Alain Moëllic, Patrick Hède, Pascal Adam. PIRIA : A general tool for indexing, search and retrieval of multimedia content. In *SPIE Electroning Imaging 2004*, San Jose, California USA, 18-22 January 2004.

[3] Y.-C. Cheng, S.-Y. Chen. Image classification using color, texture and regions. In *Image and Vision Computing*, Vol 21. Issue 9, September 2003.