Merging results from different media: Lic2m experiments at ImageCLEF 2005

Romaric Besançon, Christophe Millet
CEA-LIST/LIC2M
BP 6 92265 Fontenay-aux-Roses CEDEX - FRANCE
besanconr@zoe.cea.fr,milletc@zoe.cea.fr

Abstract

In the ImageCLEF 2005 campaign, the LIC2M participated in the ad hoc task, the medical task and the annotation task. For both ad hoc and medical task, we perform experiments on merging the results of two independent search systems: a cross-language information retrieval system exploiting the text part of the query and a content-based image retrieval system exploiting the example images given with the query. The results show that a well-tuned merging may improve performance, but the tuning is made difficult because the performance of each system highly depends on the corpus and queries. Annotation task has been performed using a KNN classifier with the image indexes of our CBIR system.

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; I.4.7 [**Computing Methodologies**]: Image Processing and Computer Vision—*Feature Measurement*

General Terms

Measurement, Performance, Experimentation

Keywords

Linguistic Processing, Cross-lingual Text Retrieval, Content Based Image Retrieval

1 Introduction

ImageCLEF campaign aims at studying cross-language image retrieval, that potentially uses text and image matching techniques. The LIC2M participated in ImageCLEF 2005 to perform experiments on merging strategies to integrate the results obtained from the cross-language text retrieval system and the content-based image retrieval system that are developed in our lab.

In both ad hoc and medical tasks of the ImageCLEF 2005 campaign, text and visual information were provided for the queries. In ad hoc task, the basic query is textual (title and narrative are provided), but two example images are provided; in medical task, query images are given and a short textual description give precisions about the research goal. We applied the same strategy for the two tasks, using our general-domain systems for multilingual text retrieval and content-based image retrieval, taking into account both text and visual part of the query and applying *a posteriori* merging strategies on the results provided independently by each system.

We present in section 2 the retrieval systems for text and image and the merging strategies used. We then present the results obtained for the ad hoc task and the medical task in sections 3 and 4 respectively. The strategy and results for the annotation task are presented in section 5.

2 Retrieval systems

2.1 Multilingual Text Retrieval System

The multilingual text retrieval system used for these experiments is basically the same as the one used for the previous CLEF campaigns, and a more detailed description can be found in [1]. The system has not been specially adapted to work on the text of the ImageCLEF corpora, and has simply been used as is. In particular, for both the ad hoc and medical 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 annotations): all fields containing some text have been taken as is. No adaptation has been made to take into account the specificities of medical texts (specialized vocabulary). Notice that this system is not only cross-lingual but multilingual, because it integrates a concept-based merging technique to merge results found in each target language. Its basic principle is briefly described here.

Document and query processing The documents and queries 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. The elements extracted from the documents are indexed into inverted files. The elements extracted from the queries are used as query "*concepts*". Each concept is reformulated into a set of *search terms* for each target language, either using a monolingual expansion dictionary (that introduces synonyms and related words), or using a bilingual dictionary.

Document Retrieval Each search term is searched in the index, and documents containing the term are retrieved. All retrieved documents are then associated with a *concept profile*, indicating the presence of query concepts in the document. This concept profile depends on the query concepts, and is language-independent (which allow merging results from different languages). Documents sharing the same concept profile are clustered together, and a weight is associated with 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-based Image Retrieval System

The content-based image retrieval system we used in ImageCLEF 2005 is the system PIRIA (Program for the Indexing and Research of Images by Affinity)[3], developed in our lab. The query image 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 signatures. 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 This indexer first quantifies the image, and then, for each quantified color, it computes how much this color is connex. It can also be described as a border/interior pixel classification [4]. The distance used for the color indexing is a classical L2 norm.

Texture Indexing   A global texture histogram is used for the texture analysis. The histogram is computed from the Local Edge Pattern descriptors [2]. These descriptors describe the local structure according to the edge image computed with a Sobel filtering. We obtain a 512-bins texture histogram, which is associated with a 64-bins color histogram where each plane of the RGB color space is quantized into 4 colors. Distances are computed with a L1 norm.

Form Indexing   The form indexer used consists of a projection of the edge image along its horizontal and vertical axes. The image is first resized in 100x100. Then, the Sobel edge image is computed and divided into four equal sized squares (up left, up right, bottom left and bottom right). Then, each 50x50 part is projected along its vertical and horizontal axes, thus giving a 400-bins histogram. The L2 distance is used to compare two histograms.

2.3   Search and Merging Strategy

For both ad hoc and medical task, the queries contain textual and visual information. Textual information is used to search relevant text documents with multilingual text retrieval system. For ad hoc task, each text document corresponds to a single image: the images corresponding to the relevant texts are then given as results. For the medical task, a text document may be associated with several images. In that case, the score obtained by the text documents is given to each image it is associated with: the first 1000 images in this image list are kept.

Since the PathoPic corpus of the medical task contains annotations in English and German that are associated with the same image, the multilingual retrieval system may return both English and German annotations as relevant documents (maybe with different scores), creating duplicate elements in the result list. In this case, the score associated with the corresponding image is the best score returned. To make sure that the number of retrieved images is 1000, we set the number of retrieved documents for the text retrieval system at 2000 for the medical task[1].

Independently, visual information was used by the CBIR system to retrieve similar images. Queries contain several images: a first merging has been performed to obtain a single image list from the results of each query image: the score associated to result images is set to the max of the scores obtained for each query image.

Merging the results obtained by each system is simply done by a weighted sum of the scores obtained by each system. To be comparable, the scores of each system are normalized, for each query, by the highest score obtained for the query. This merging is parameterized by a merging coefficient α: for a query q and an image document retrieved for this query $d \in Ret(q)$, the merging score is

$$s(d) = \alpha \times \frac{s_T(d)}{\max\limits_{d \in Ret_T(q)} s_T(d)} + (1 - \alpha) \times \frac{s_I(d)}{\max\limits_{d \in Ret_I(q)} s_I(d)}$$

where $s_T(d)$ is the score of the text retrieval system and $s_I(d)$ the score of the image retrieval system.

A conservative merging strategy has also been tested: by *conservative*, we mean that we use the results obtained by one system only to reorder the results obtained by the other (results can be added at end of list if the number of documents retrieved by main system is less than 1000). The score of a document is modified using the same merging coefficient. For example, if the merging is conservative with the text results:

$$s'(d) = \begin{cases} s(d) & \text{if } s_T(d) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

The results we obtained in ImageCLEF 2004 tend to show that this kind of conservative merging strategies gives good performances. We will use the term of *expansionist* merging strategy to denote standard merging strategy, as opposed to the conservative one.

---

[1]this duplication of results was not detected before the submission of the runs, but the technique we used for merging text and image results remove the duplicate documents.

3   Results for the Ad hoc task

In the ad hoc task, we used textual queries in English, French and Spanish. We tried using the title only (T) or the title and the narrative (T+N). Comparative results for textual retrieval only, using either T or T+N are given in Table 1. These results show that average precision is better when using the title only, but the number of relevant documents is generally better when using also the narrative part (except for French, for which it is a bit worse). This can be explained by the fact that narrative introduce more words that allow to increase the total number of documents retrieved (for English and Spanish, there are 6 queries for which the system does not find 1000 documents matching title only, only 3 for French), and the number of relevant documents. But narrative also introduces more noise, which makes the precision decrease.

	eng		fre		spa	
	T	T+N	T	T+N	T	T+N
map	**0.246**	0.224	**0.186**	0.146	**0.191**	0.151
relret	1246	**1401**	**1237**	1184	1085	**1153**
r1000	65%	**73.1%**	**64.6%**	61.8%	56.6%	**60.2%**

Table 1: Ad hoc task: comparative results using title or title+narrative: mean average precision (*map*), number of relevant documents (*relret*) and recall at 1000 documents (*r1000*)

We present in Table 2 the results obtained by the merging of the two systems, using the texture indexer for the CBIR system. The results are presented for conservative and expansionist strategies and for different values of the merging coefficient α (when $\alpha = 1$, the search is only based on text, when $\alpha = 0$, the search is only based on images). Values below 0.5 are not presented but does not give better results. For expansionist strategy, the results are given for the mean average precision (*map*) and the number of relevant documents retrieved (*relret*); for conservative strategy, only the *map* is presented (*relret* is constant).

These results show that this simple merging of text and image results based on a weighted sum of the scores can increase the mean average precision (gain of 17 or 18%) and the best value for α is around 0.7 (though differences with surrounding values are small).

Concerning conservative/expansionist strategies, our previous experiments in ImageCLEF showed that the StAndrews collection, composed of old photographs, is not well adapted the kind of image indexers we use, that rely mostly on color for segmentation. We therefore chose the text retrieval as base for conservative merging. Looking at the relevant documents retrieved prove us right: text retrieval allow to retrieve 1246 relevant documents, whereas image retrieval only retrieve 367 relevant documents (239 of which were also found by the text retrieval system). However, the two merging strategies give comparable results, even though, as one can expect, the performance of the expansionist strategy decreases faster with α.

conservative merging (map)				expansionist merging (map / relret)			
α	eng	fre	spa	α	eng	fre	spa
1	0.246	0.186	0.191	1	0.246/ 1246	0.186 / 1237	0.191 / 1085
0.9	0.274	0.214	0.209	0.9	0.274/**1254**	0.214 / 1254	0.208 / 1088
0.8	**0.282**	**0.221**	0.212	0.8	**0.28** / 1221	**0.221**/ 1251	0.212 / 1091
0.7	0.28	0.22	0.223	0.7	0.275/ 1214	0.204 /**1254**	**0.222**/ 1094
0.6	**0.282**	0.218	0.225	0.6	0.26 / 1166	0.176 / 1231	0.219 / 1086
0.5	0.276	0.213	**0.227**	0.5	0.236/ 1143	0.162 / 1146	0.217 /**1098**
0	0.068	0.068	0.068	0	0.068/ 367	0.068 / 367	0.068 / 367

Table 2: Ad hoc task: comparative results for the merging strategies using texture indexer for the image retrieval

Similar results are presented in Table 3 using the color indexer for the CBIR system. Results are comparable: for this corpus, the two image indexers tend to retrieve similar documents (2/3

of relevant documents retrieved by both systems are identical).

	conservative merging (map)				expansionist merging (map / relret)		
α	eng	fre	spa	α	eng	fre	spa
1	0.246	0.186	0.191	1	0.246/ 1246	0.186 / 1237	0.191 / 1085
0.9	0.274	0.214	0.207	0.9	0.274/**1250**	0.214 / 1244	0.207 / 1096
0.8	0.281	**0.219**	0.21	0.8	**0.28** / 1211	**0.218**/ 1241	0.21 / 1096
0.7	0.281	0.216	0.222	0.7	0.273/ 1196	0.191 /**1269**	0.22 / 1104
0.6	0.282	0.209	**0.226**	0.6	0.255/ 1148	0.167 / 1250	**0.221**/**1131**
0.5	**0.283**	0.208	**0.226**	0.5	0.238/ 1110	0.158 / 1128	0.217 / 1124
0	0.065	0.065	0.065	0	0.065/ 330	0.065 / 330	0.065 / 330

Table 3: Ad hoc task: comparative results for the merging strategies using color indexer for the image retrieval

Submitted runs for the ad hoc task in the ImageCLEF 2005 campaign were, for English, French and Spanish, text only (T and T+N), plus a conservative merging of text results based on title only and image results based on the texture indexer, with $\alpha = 0.9$.

4   Results for the Medical task

In the medical task, we tested text retrieval using queries in English, French and German (searching for each in all target languages).

Based on our experiments in ImageCLEF 2004, we assumed that image retrieval for the medical task gives good results. Submitted runs for the medical task in the ImageCLEF 2005 campaign include runs based on visual queries only (texture and color indexers), and for English, French and German, a conservative merging of image results based on the texture indexer and text results, with $\alpha = 0.9$. Unfortunately, the use of texture or color indexer with the ImageCLEFmed 2005 visual queries gave poor results, and conservative merging based on these results does not give much better results[2].

We present in Table 4 the results obtained by the merging of text and image systems, using the texture indexer for the CBIR system, with different values of the merging coefficient α, and for conservative and expansionist merging strategies (conservative strategy based on text results).

Except for German (for which our linguistic processing is clearly not well adapted to medical text), the conservative merging strategy improves performances (the best merging coefficient seems to be around 0.5). Expansionist merging gives comparable results: improvement of mean average precision is less important, but the number of relevant documents retrieved is generally improved, which tends to prove that both systems retrieve different documents[3]: conservative merging improves the ordering of documents retrieved by one system whereas expansionist merging improves the number of documents retrieved.

We present in Table 5 similar results using the color indexer for visual retrieval. Results are slightly worse, but the same kind of tendencies as for the texture indexer can be noticed.

5   Annotation task

For the automatic annotation task, we submitted three runs, each corresponding to one of the three indexers described in section 2 (Color, Texture and Form).

All images are first indexed with the chosen indexer. Then, a k-Nearest Neighbor classifier is used to classify the indexed images. Odd numbers from 3 to 13 have been tested for k for each

---

[2]Furthermore, we detected a bug in submitted runs, concerning the document identifier matching (*1* vs. *0000001*) that made the Peir corpus documents ignored in text retrieval results.

[3]We verified that text results with English queries contain 999 relevant images, image results with texture indexer contain 822 relevant images and only 218 images were common to the two systems.

conservative merging (map)				expansionist merging (map / relret)			
α	eng	fre	ger	α	eng	fre	ger
1	0.0843	0.0899	0.0179	1	0.0843 / 999	0.0899/ 1059	0.0179 / 466
0.9	0.11	0.124	0.0237	0.9	0.11 / 1122	0.124 / 1236	0.0238 / 501
0.8	0.11	0.127	0.0278	0.8	0.11 / 1133	**0.129**/ 1275	0.0278 / 501
0.7	0.115	0.129	0.031	0.7	0.114 / 1153	**0.129**/**1325**	0.0312 / 507
0.6	**0.122**	0.131	0.037	0.6	**0.118**/ 1161	0.127 / 1334	0.0388 / 571
0.5	**0.122**	**0.135**	0.0411	0.5	0.108 /**1192**	0.123 / 1281	**0.0451**/**660**
0	0.0465	0.0465	**0.0465**	0	0.0307/ 643	0.0307/ 643	0.0307 / 643

Table 4: Medical task: comparative results for the merging strategies using texture indexer for the image retrieval

conservative merging (map)				expansionist merging (map / relret)			
α	eng	fre	ger	α	eng	fre	ger
1	0.0843	0.0899	0.0179	1	0.0843 / 999	0.0899/ 1059	0.0179 / 466
0.9	0.0932	0.116	0.0209	0.9	0.0932 / 1112	0.115 / 1226	0.021 / 499
0.8	0.0933	0.119	0.0234	0.8	0.0929 / 1112	0.12 / 1232	0.0235 / 499
0.7	0.096	0.12	0.0278	0.7	0.0917 /**1115**	**0.117**/**1248**	0.028 / 507
0.6	0.103	0.121	0.0337	0.6	**0.0954**/ 1102	0.111 / 1246	0.0331 / 538
0.5	**0.106**	**0.122**	**0.0367**	0.5	0.0908 / 1074	0.105 / 1153	**0.0356**/ 580
0	0.0307	0.0307	0.0307	0	0.0307 / 643	0.0307/ 643	0.0307 /**643**

Table 5: Medical task: comparative results for the merging strategies using color indexer for the image retrieval

indexer, and evaluated with the leave-one-out method. The best k were 3 for the form indexer and 9 for the color indexer and the texture indexer.

The attributed class is decided by a majority vote of the nearest neighbors. In case of ties, distances to nearest neighbors are used (for example, in 9-NN, if 4 neighbors are from a class A, 4 neighbors from a class B, and 1 from another class, we use the distances between the requested image and its neighbors to select the nearest class).

We present in Table 6 the results obtained for each of the indexers. It is not a surprise that the form indexer performed better than the others, as all the images in the database were in grey levels, and the form indexer is designed for such images, whereas the color and texture indexers are not well adapted to it (remember that the texture indexer includes a 64-bins color histogram).

	9-NN Color	9-NN Texture	3-NN Form
error rate	46.0 %	42.5 %	**36.9 %**

Table 6: Results for the automatic annotation task

6   Conclusion

The experiments performed by the LIC2M in the ImageCLEF 2005 campaign show that merging results from different media may increase the performance of a search system: a well-tuned *a posteriori* merging of the results obtained by two general purpose systems (no particular adaptation of the systems was made for the two tasks) can improve the mean average precision by at least 15%.

The difficulty relies on the tuning of the merging strategy. We used a simple weighted sum of the scores given by each system but the importance given to each system should rely on the performance of the system on a particular corpus, that is not easily predicted (best strategy for the

ImageCLEF 2004 medical task appears to be opposite to the best strategy for ImageCLEF 2005 medical task, that has a more varied corpus and more difficult visual queries).

Further experiments will be undertaken to try to make the systems give a confidence score associated with its results and adapt the merging strategy according to this confidence. Other more sophisticated merging strategies will also be considered.

References

[1] Romaric Besançon, Gaël de Chalendar, Olivier Ferret, Christian Fluhr, Olivier Mesnard, and Hubert Naets. Concept-based searching and merging for multilingual information retrieval: First experiments at clef 2003. In Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck, editors, *Comparative Evaluation of Multilingual Information Access Systems*, pages 174–184. Springer, 2004.

[2] Ya-Chun Cheng and Shu-Yuan Chen. Image classification using color, texture and regions. *Image and Vision Computing*, 21(9), September 2003.

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

[4] Renato. O. Stehling, Mario. A. Nascimento, and Alexandre X. Falcão. A compact and efficient image retrieval approach based on border/interior pixel classification. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, McLean, Virginia, USA, 2002.