UNED-UV at Medical Retrieval Task of ImageCLEF 2011

A.Castellanos[1], X. Benavent[2], J.Benavent[2], Ana García-Serrano[1]

[1] Universidad Nacional de Educación a Distancia, UNED

[2] Universitat de València

xaro.benavent@uv.es,agarcia@lsi.uned.es

Abstract. The main goal of this paper it is to present our experiments in ImageCLEF 2011 Campaign (Medical Retrieval Task). This edition we use textual and visual information, based on the assumption that the textual module better captures the meaning of a topic. So that, the TBIR module works firstly and acts as a filter, and the CBIR system reorder the textual result list. We also investigate if query expansion with image terms or with modality classification could be a way to improve base queries This paper is profiting on the work done in previous years on ImageCLEF (Wikipedia Retrieval Task). In this edition we submitted a total of ten runs (4 textual and 6 mixed). Textual ones have better results (being two of them 2nd and 6th within their category). Mixed runs are about the middle of the results, although have demonstrated that can improve the only textual results. With our results we have proved that query expansion with term concerning image type of the query is a promising way to further research.

Keywords: Query Expansion, Textual-based Retrieval, Content-based Retrieval, Merging

1 Introduction

The main goal of this paper it is to present our experiments in ImageCLEF 2011 Campaign (Medical Image retrieval task) [1]. In this working note, we rather focus on our participation in two sub-tasks of the Medical Retrieval Tasks (Image Modality Classification and Ad-hoc Image Retrieval).

This ImageCLEF edition our group presents a way of working using the information of the Content Based Image Retrieval (CBIR) system and the information of the Textual Based Image Retrieval (TBIR) system. We use the work done in this regards in previous editions of ImageCLEF [2], based on the assumption that the conceptual meaning of a topic is initially better captured by the text module itself than by the visual module, so our merging method gives greater weight.

In this field numerous approaches have been researched in previous works [3] [4], our view raises a different type of merging, the TBIR system works firstly over the whole database working as a filter, and then the Valencia University CBIR system reorders the filtered textual result list, taking into account the visual features.

Concerning the TBIR subsystem, we have worked on two different approaches. The first of them pretended to check if the query expansion with term relatives to image type could achieve a significant improvement of the results obtained. In past years of ImageCLEF, different methods of query expansion have been posed. The bulk of them were focused on using external medical sources like MESH [5] or UMLS [6], also has been raised the expansion of queries with another external sources like Wikipedia [7]. Our aim is to prove another type of query expansion that can complement the studies presented above.

The second approach was focused on investigating how vary the results returned by the queries adding the information relating to the classification of the images provided by ImageCLEF.

The CBIR subsystem is quite similar to the system used in previous works of ImageCLEF [2]. The CBIR subsystem uses its own low-level features or the CEDD features [11], depending on the experiment in order to test the influence of the low-level features in the final result. Two different algorithms have also been used: logistic regression relevance feedback algorithm and an automatic algorithm with the Tanimoto distance.

A more detailed presentation of the system, the submitted experiments, and the obtained results is included in the following sections.

2 System Description

The global system includes two main subsystems: the TBIR and the CBIR (Fig.1.). The TBIR subsystem is responsible for preprocessing and indexing the textual information both of the images, textual annotations of images and queries.

The TBIR subsystem acts over the whole images of the database as a filter. Only the images returned by TBIR module are sending to the CBIR system. In a second step, the CBIR system works over the set of filtered images reordering this list taking into account the visual information of the image and the score given by TBIR module to each image.

2.1 Text-based Index and Retrieval

This module (TBIR) is in charge of the textual image retrieval using the text associated with each image in the collection.

Previously, to be able to work with the collection, this was preprocessed. Later, it has been carried out the indexing of the images and their subsequent search and retrieval of images for each query through Solr[1], a search platform from Lucene[2] project. The result of this process is an image list that is ranked according to their similarity with the corresponding query, in accordance only with the textual information. Below, is explained in more detail each of the different stages performed by TBIR module:

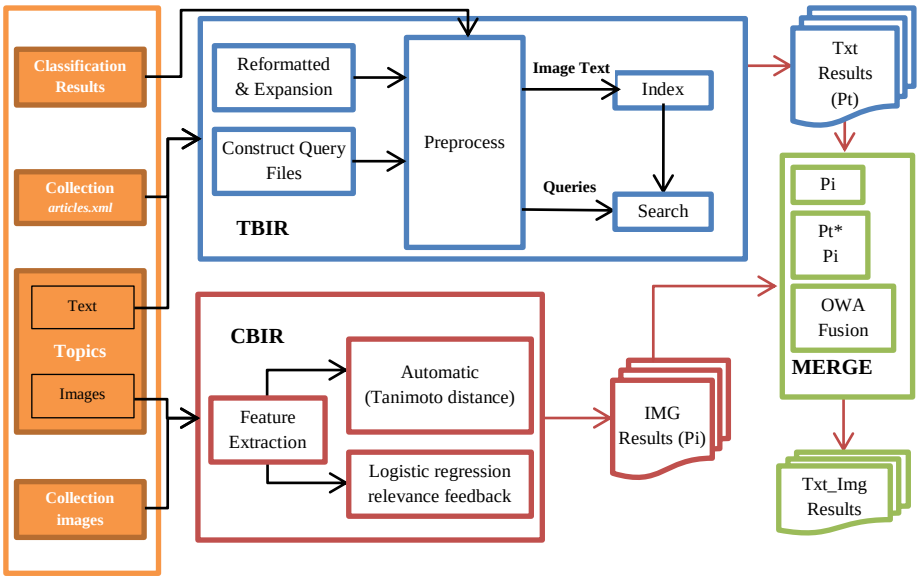

Fig. 1. - System Overview

Collection Reformatted. The whole collection is provided in a single xml file. To facilitate the work, we split the file, creating a new xml file for each of the articles in which the collection is divided, resulting in 55634 xml files. It is not necessary to extract the textual information from xml files, as Solr supports xml as input files.

[1] http://lucene.apache.org/solr/

[2] http://lucene.apache.org/java/docs/index.html

Collection Expansion: In order to investigate how the image classification affects the results of queries, it has decided to include an additional field (called tags) in the xml file descriptors of the images. This field stores the modality of the image established by the classification provided by ImageCLEF.

Construct Queries File. Based on original query file, four query files are constructed in order to address the two approaches we want to investigate: query expansion and inclusion of modality classification. Each of these files corresponds to one of the four runs explained below

Preprocess. Textual information is preprocessed in different ways in order to improve search and retrieval. The order in which transformations are applied is as follows: 1) special characters deletion: characters with no statistical meaning, like punctuation marks or blanks, are eliminated; 2) stopwords detection: deletion of semantic empty words in English language, 3) stemming: reduction of word to their base form and 4) convert all words to lower case.

Indexing. The indexing is done automatically by Solr, this requires to configure which fields must have the index and create a handler that establish how Solr have to read the xml file input and add their information to the index. The indexing itself is made by Lucene. The time of indexing, about 7 minutes, is relatively short considering the size of the collection.

Searching. This process is started manually, running each query by a Solr interface. The results, returned in xml format, are transformed to the trec-eval format, in order to merge these textual results with visual results and check the results using the UV tool.

2.2 Content-Based Information and Visual Retrieval

The VISION-Team at the Computer Science Department of the University of Valencia has its own CBIR system, and that has been used in previous ImageCLEF editions (Photo-retrieval task [8]). Last edition, the focus of the work was the testing of three different visual algorithms applied to the results retrieved by the textual module: the automatic, the relevance feedback and the query expansion, obtaining the best results with the relevance feedback algorithm. Therefore, this edition we have used the relevance feedback algorithm and our work focus on testing the behavior of our own low-level features with the low-level features given by the organization (the CEDD algorithm described in (1)).

Extraction of low level features:

The first step at the Visual Retrieval system is extracting these features for all the images on the database as for each of the cluster query topic images for each question. Instead of using the low-level features provided by the organization, we have used our own features:

- **Color information:** Color information has been extracted calculating both local and global histograms of the images using 10x3 bins on the HS color system. Local histograms have been calculated dividing the images in four fragments of the same size. A bidimensional HS histogram with 10x3 bins is computed for each patch. Therefore, a feature vector of 222 components represents the color information of the image.
- **Texture information:** Two types of texture features are computed: The granulometric distribution function, using the coefficients that result of fitting the distribution function with a B-spline basis. And, the Spatial Size Distribution. We have used two different versions of it by using as the structuring elements for the morphological operation that get size both a horizontal and a vertical segment.

Automatic algorithm.

This is the most common algorithm in a CBIR system. Each image in the database has an associated low level feature vector. Concretely, we have used for this algorithm the low level features given by the organization (CEDD).

The next step is to calculate the similarity measurement between the feature vectors of each image on the database and the N query images. The distance metric applied in our experiments is the Tanimoto

As we have N query images, we will obtain N visual result lists, one for each query image in the topic. These N result lists are merged by using an average OWA operator.

Relevance feedback algorithm based on logistic regression.

This algorithm works differently to the two previous ones. Therefore, we will explain the concept of relevance feedback and the adjustments made to get a good performance of the algorithm for the proposed tasks [9]. Relevance feedback is a term used to describe the actions performed by an user to interactively improve the results

of a query by reformulating it. An initial query formulated by a user may not fully capture his/her wishes. Users then typically change the query manually and re-execute the search until they are satisfied. By using relevance feedback, the system learns a new query that better captures the user's need for information. The user enters his/her preferences every iteration through the selection of relevant and non-relevant images.

We will explain the way the logistic regression relevance feedback algorithm works. Let us consider the (random) variable Y giving the user evaluation where Y=1 means that the image is positively evaluated and Y=0 means a negative evaluation. Each image in the database has been previously described by using low-level features in such a way that the j-th image has the k-dimensional feature vector xj associated. Our data will consist of (xj, yj), with j=1,…,n, where n is the total number of images, xj is the feature vector and yj the user evaluation (1=positive and 0=negative). The image feature vector x is known for any image and we intend to predict the associated value of Y. In this work, we have used a logistic regression where P(Y=1|x) i.e. the probability that Y=1 (the user evaluates the image positively) given the feature vector x, is related with the systematic part of the model (a linear combination of the feature vector) by means of the logit function. For a binary response variable Y and p explanatory variables X1,…,Xp, the model for $\pi(x)=P(Y=1|x)$ at values x=(x1,…,xp) of predictors is $\text{logit}[\pi(x)]=\alpha+\beta 1x1+\ldots+\beta pxp$, where $\text{logit}[\pi(x)]=\ln(\pi(x)/(1-\pi(x)))$. The model parameters are obtained by maximizing the likelihood function given by:

$$l(\beta)=\prod^{n}\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \quad \textbf{(1)}$$

The maximum likelihood estimators (MLE) of the parameter vector β are calculated by using an iterative method.

We have a major difficulty when having to adjust a global regression model in which we take the whole set of variables into account, because the number of selected images (the number of positive plus negative images) is typically smaller than the number of characteristics. In this case, the regression model adjusted has as many parameters as the number of data and many relevant variables could be not considered. In order to solve this problem, our proposal is to adjust different smaller regression models: each model considers only a subset of variables consisting of semantically related characteristics of the image. Consequently, each sub-model will associate a different relevance probability to a given image x, and we face the question of how to combine them in order to rank the database according to the user's preferences. This problem has been solved by means of an ordered averaged weighted operator (OWA) [10]

In our case, we have adapted the manual relevance feedback to an automatic performance. The examples and the counter-examples (positive and negative images) are automatically selected for each topic. The examples are the query images of the topic plus N images taken from the first positions of the textual result list. The counter-examples are the M latest positions of the textual result list. The relevance feedback algorithm is executed once.

2.3 Merging Results

We use a merging process for combining textual and visual results in order to try to improve them. To perform this merging process we use TBIR system like filter over the whole collection, only the results retrieved by TBIR system are passed through the CBIR system.

The CBIR system search taking into account visual characteristics of the images contained in each query. Based on the score of the results retrieved by this search, the images returned by TBIR system are reordered in three different ways (taking into account only score of the visual results, taking into account both visual and textual results and by combining textual and visual through an OWA fusion method).

3 Experiments

We have participated in ad-hoc image-based retrieval task of Medical Image Retrieval Task 2011. Finally, we submitted 10 runs: 4 textual and 6 mixed (textual and visual). This runs is shown in **Table 1.**

With this runs we intend to study how the two approaches raised for this work affect to the results. For the first approach, query expansion about image terms, we present four different types of queries for textual retrieval that are explained in more detail below. Besides this in two of this runs (UNED-UV_05 and UNED-UV_06), we cover the other method, including modality classification. Hoping to improve results, we present another six runs merging textual retrieval with content-based retrieval

To do this we present 4 textual runs, based on 4 different types of queries explained in previous sections, UNED-UV_02 like baseline, UNED-UV_03 using query expansion and UNED-UV_05 and UNED-UV_06 including modality classification.

The mixed runs are based on UNED-UV_05 and UNED-UV_06 (which it expected would have better results). The results obtained with UNED-UV_05 it have been passed through CBIR module in three different ways: 1) reordering the results taking into account only the weight given by CBIR module (Pi) [UNED-UV_11], 2) reordering based on the product of textual and visual weights (Pt*Pi) [UNED-UV_17]

and 3) reordering combining textual and visual results by an OWA weighted with 0.7 and 0.3 respectively [UNED-UV_23]. For the UNED-UV_06, following the same structure we obtained runs [UNED-UV_12], [UNED-UV_18] and [UNED-UV_24]

3.1 Detailed Description of Selected Runs

UNED_UV_02_TXT_AUTO_EN: Baseline Run. Terms concerning to the image type (image, photograph …) are deleted in order to focus on medical terms included in each query. It's expected to improve results by reducing the noise introduced by these terms. We do not remove terms that refer directly to a type of modality classification (CT, XR …).

Original query: *photographs of benign or malignant skin lesions*
Expanded query: *benign OR malignant skin lesions*

UNED_UV_03_TXT_AUTO_EN: Instead of deleting image terms, these are replaced by modality classification terms, whenever possible (x-ray = XR).

Original query: *all x-ray images containing one or more fractures*
Expanded query: *all (XR OR "x-ray") AND (one or more fractures)*

UNED_UV_05_TXT_AUTO_EN: In this RUN we use the classification provided by ImageCLEF. It's intended to demonstrate that including information concerning to the classification of images significantly improves the results returned.

For this run, were added to each query another query against tag field (included in collection expansion) with the value of classification modality expected for the original query. The original query is weighted with 0.3 and query against tag field is weighted with 0.7. With these weights we want to avoid bias of the results, caused by medical terms.

Original query: *chest CT images with emphysema*.
Expanded query: *(tags:CT)^0.7 OR (chest CT AND emphysema)^0.3*

UNED_UV_06_TXT_AUTO_EN: This RUN is similar to previous. We also included classification of images and expansion of the original queries with another query against tag field, but in this case, instead of aggregate this second query, it's used for filtering the results returned by the original query.

Original query: chest CT images with emphysema.
Expanded query: chest CT AND emphysema→ results → tags:CT → filtered results

UNED_UV_11_TXTIMG_AUTO_EN: In this RUN we used the results returned by UNED_UV_05_TXT_AUTO_EN as input data and are reordered taking into account the weights established by CBIR module.

UNED_UV_12_TXTIMG_AUTO_EN: Same as above RUN, but using UNED_UV_06_TXT_AUTO_EN instead of UNED_UV_05_TXT_AUTO_EN.

UNED_UV_17_TXTIMG_AUTO_EN: UNED_UV_05_TXT_AUTO_EN results are used like input and are reordered based on the products of weights of TBIR and CBIR modules.

UNED_UV_18_TXTIMG_AUTO_EN: Same as above but using UNED_UV_06_TXT_AUTO_EN results like input.

UNED_UV_23_TXTIMG_AUTO_EN: UNED_UV_05_TXT_AUTO_EN results are used like input and are reordered according to an OWA with a weight of 0.3 for CBIR weighting and 0.7 for TBIR weighting.

UNED_UV_24_TXTIMG_AUTO_EN: Same as above but using UNED_UV_06_TXT_AUTO_EN results like input.

Table 1. - Submitted textual and mixed experiments

		CBIR		TBIR
Run	**Modality**	**Image**	**Merge**	**Method**
UNED_UV_02_TXT_AUTO_EN	Text	-	-	Baseline
UNED_UV_03_TXT_AUTO_EN	Text	-	-	Query Expansion
UNED_UV_05_TXT_AUTO_EN	Text	-	-	Include Modality Classification - Aggregation
UNED_UV_06_TXT_AUTO_EN	Text	-	-	Include Modality Classification - Filtering
UNED_UV_11_TXTIMG_AUTO_EN	Mixed	LR_RF	Pi	-
UNED_UV_12_TXTIMG_AUTO_EN	Mixed	LR_RF	Pi	-
UNED_UV_17_TXTIMG_AUTO_EN	Mixed	LR_RF	Pt*Pi	-
UNED_UV_18_TXTIMG_AUTO_EN	Mixed	LR_RF	Pt*Pi	-
UNED_UV_23_TXTIMG_AUTO_EN	Mixed	LR_RF	OWA(Orness(0.3))	-
UNED_UV_24_TXTIMG_AUTO_EN	Mixed	LR_RF	OWA(Orness(0.3))	-

4 Results

Our results for the submitted experiments are shown in Table 2. As reflected in the table, concerning only mixed experiments, our best results are those based on Run6 (Run6_TxtImgOwaOr03, at the 19th position and Run6_TxtImg_PtPi at the 20th

position of a total of 40). For the textual modality our results are considerably better (Run2_Txt at the 2nd and Run3_Txt at the 6th position, at the 10% first results).

Table 2 -Results for the submitted experiments

Pos	Run	Mode	Map	P@10	P@20	R-prec	B-prec.	Num_rel_ret
19	Run6_TxtImg_OwaOr03	Mixed	0.1346	0.3467	0.2867	0.1666	0.1604	565
20	Run6_TxtImg_PtPi	Mixed	0.1311	0.3333	0.2817	0.1651	0.1557	565
21	Run5_TxtImg_OwaOr03	Mixed	0.1299	0.3200	0.2567	0.1613	0.1641	710
23	Run5_TxtImg_PtPi	Mixed	0.1176	0.2800	0.2100	0.1575	0.1614	705
30	Run6_TxtImg_Pi	Mixed	0.0891	0.2400	0.1933	0.1206	0.1288	508
33	Run5_TxtImg_Pi	Mixed	0.0699	0.1667	0.1517	0.1017	0.1394	755
2	Run2_Txt	Textual	0.2158	0.3533	0.3383	0.2470	0.2514	1383
6	Run3_Txt	Textual	0.2125	0.3867	0.3317	0.2468	0.2430	1138
53	Run6_Txt	Textual	0.1309	0.3433	0.2733	0.1652	0.1597	564
55	Run5_Txt	Textual	0.1270	0.3100	0.2517	0.1565	0.1622	651

With the textual experiments we aimed to analyze two approaches this year. First of all we wanted investigate if query expansion, with terms relatives to the image type, improves the retrieved results. As it can be viewed in the table, runs in this sense (Run2_Txt and Run3_Txt) provide excellent results. The other approach that we pretend investigate was the inclusion of the images classification provided by the organization. In this regard we submitted two runs (Run6_Txt and Run5_Txt). The results of these runs are not good as we expected, worsen the results of the baseline runs.

In order to improve the textual results, we submitted 6 mixed runs, combining textual and visual retrieval. As is evident by the results, mixed results slightly improve textual results (especially using an OWA like fusion method). Since we expected that better results was provided by Run6_Txt and Run5_Txt, we based our 6 experiments on these, so mixed results are not as good as could be if we had used Run2_Txt or Run3_Txt.

5 Concluding Remarks and Future Work

In this year we have participated by first time in Medical Image Retrieval Task and yet our runs have had very satisfactory results. Our bests results are for the textual modality, two of our runs are in the 2nd and 6th position respectively in the textual category. In addition, most of our runs are between first 50% in their category.

Concerning to our approaches, we have demonstrated that query expansion with terms relatives to image concepts improves in a significantly way the retrieved results and show much more effective than the query expansion with modality classification of the images.

Can be gathered from our mixed runs is that the combination and textual and visual results can improve results, although in a slightly way. In this regard we pretend to continue investigating in order to refine the process of fusion textual and visual results.

Future work pretends to go beyond in the work of query expansion conducted in this work, including external sources like MESH [5] that have demonstrated that significantly improves the results for this task, as well as further research in the mixed (textual and visual) retrieval.

Acknowledgments. This work has been partially supported for Regional Government of Madrid under Research Network MA2VIRMR (S2009/TIC-1542), for Spanish Government by project BUSCAMEDIA (CEN-20091026) and by project MCYT TEC2009-12980.

References

1. Jayashree Kalpathy-Cramer, Henning Müller, Steven Bedrick, Ivan Eggel, Alba Garcia Seco de Herrera, Theodora Tsikrika, "The CLEF 2011 medical image retrieval and classification tasks", CLEF 2011 working notes,Amsterdam, The Netherlands, 2011.
2. Benavent, J, et al: Experiences at ImageCLEF 2010 using CBIR and TBIR mixing information aproaches. Conference on Multilingual and Multimodal Information Access Evaluation CLEF 2010. Working Notes for the CLEF, Padua (Italy), 2010.
3. Chevallet, Jean-Pierre and Lim, Joo-Hwee: Using Ontology Dimensions and Negative Expansion to solve Precise Queries in CLEF Medical Task. 2005.
4. Torjmen, Mouna, Pinel-Sauvagnat, Karen and Boughanem, Mohand. s.l. : Methods for combining content-based and textual-based approaches in medical image retrieval. Evaluating Systems for Multilingual and Multimodal Information Access, 2009.
5. Bedrick, Steven and Kalpathy-Cramer, Jayashree: Improving Retrieval Using External Annotations: OHSU at ImageCLEF 2010. 2010.

6. Villena-Román, Julio, Lana-Serrano, Sara and González-Cristobal, José Carlos. s.l. : MIRACLE at ImageCLEFmed 2007: Merging Textual and Visual Strategies to Improve Medical Image Retrieval. Advances in Multilingual and Multimodal Information Retrieval, 2007.

7. Clinchant, Stéphane, et al.: XRCE's Participation in Wikipedia Retrieval, Medical Image Modality Classification Ad-hoc Retrieval Tasks of ImageCLEF 2010. Conference on Multilingual and Multimodal Information Access Evaluation CLEF 2010. Working Notes for the CLEF, Padua (Italy), 2010.

8. Granados, Ruben, et al :MIRACLE (FI) at ImageCLEFphoto 2009. Cross-Languaje Evaluation Forum CLEF 2009. Working Notes for the CLEF, Corfu(Grecia) 2009.

9. Leon, T, et al.: Applying logistic regression to relevance feedback in image retrieval, Pattern Recognition. 2007, Vol. 40, pp. 2621-2632.

10. Yaguer, R. On ordered weighted averaging aggregation operators in multi criteria decision making. IEEE Transactions Systems Man and Cybernetics. 1998, Vol. 18, pp. 183-190.

11. S. A. Chatzichristofis, K. Zagoris, Y. S. Boutalis and N. Papamarkos, "Accurate Image Retrieval based on Compact Composite Descriptors and Relevance Feedback Information", International Journal of Pattern Recognition and Artificial Intelligence (IJPRAI), Volume 24, Number 2 / February, 2010, pp. 207-244, World Scientific.