MIRACLE's Naive Approach to Medical Images Annotation

Julio Villena-Román[1,3], José Carlos González-Cristóbal[2, 3]
José Miguel Goñi-Menoyo[2], José Luís Martínez-Fernandez[1, 3]

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

jvillena@daedalus.es, jgonzalez@dit.upm.es,
josemiguel.goni@upm.es, jmartinez@daedalus.es

Abstract

One of the proposed tasks of the ImageCLEF 2005 campaign has been an Automatic Annotation Task. The objective is to provide the classification of a given set of 1,000 previously unseen medical (radiological) images according to 57 predefined categories covering different medical pathologies. 9,000 classified training images are given which can be used in any way to train a classifier. The Automatic Annotation task uses no textual information, but image-content information only. This paper describes our participation in the automatic annotation task of ImageCLEF 2005.

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, medical images, image annotation, learning algorithms, decision table, nearest-neighbour, Weka.

1 Introduction

ImageCLEF is the cross-language image retrieval track which was established in 2003 as part of the Cross Language Evaluation Forum (CLEF), a benchmarking event for multilingual information retrieval held annually since 2000. Originally, ImageCLEF focused specifically on evaluating the retrieval of relevant images of the collection using different query languages, therefore having to deal with monolingual and bilingual image retrieval (multilingual retrieval is not possible as the document collection is only in one language). Later, the scope of ImageCLEF widened and goals evolved to investigate the effectiveness of combining text and image for retrieval (text and content-based), collect and provide resources for benchmarking image retrieval systems and promote the exchange of ideas which will lead to improvements in the performance of retrieval systems in general.

In addition to the retrieval experiments, the 2005 campaign also included a new completely different task: an Automatic Annotation task. The objective is to provide the classification of a given set of 1,000 previously unseen medical (radiological) images according to 57 predefined categories covering different medical pathologies. 9,000 classified training images are given which can be used in any way to train a classifier. The Automatic Annotation task uses no textual information, but image-content information only.

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 [11],[10],[8],[3],[2],[7]. As well as bilingual, monolingual and cross lingual tasks, the team has participated in the ImageCLEF, Q&A, WebCLEF and GeoCLEF tracks.

This paper describes our participation in the automatic annotation task of ImageCLEF 2005. Although this task is clearly aimed at image analysis research groups and our areas of expertise don't include image analysis research, we decided to participate in this task adopting a naive approach which consists on isolating ourselves from the content-based analysis by using a publicly available content-based image retrieval system (GIFT [1]) and applying learning (mainly classification) techniques on the results. The main objective behind our effort to participate is to promote and encourage multidisciplinary participation in all aspects of information retrieval, no matter if it is text or content based.

2 Task goals

Automatic image classification or image annotation is an important step when searching for images from a database, as a way to limit the number of results or filter them to increase precision or as a starting point for a guided search.

In the specific context of medical images, the automatic image annotation may be used as part of a diagnosis support system [4]. This system ought to classify and register medical images, using methods of pattern recognition and structural analysis to describe the image content in a feature based, formal and generalized way. The formalized and normalized description of the images then would be used as a mean to compare images in the archive which allows a fast and reliable retrieval. In addition to the queries on an existing electronic archive, the automatic classification allows a simple insertion of conventional radiographs into the system without interaction and therefore costly editing of diagnostic findings.

Based on the IRMA (Image Retrieval in Medical Applications) project [6], a database of 9,000 fully classified radiographs taken randomly from medical routine is made available and can be used to train a classification system. 1,000 radiographs for which classification labels are not available to the participants have to be classified, which is the objective of the Automatic Annotation task in ImageCLEF 2005.

The aim is to find out how well current techniques can identify image modality, body orientation, body region, and biological system examined based on the images. The results of the classification step can be used for multilingual image annotations as well as for DICOM (Digital Imaging and Communications in Medicine) header corrections.

The images are annotated with complete IRMA code, a multi-axial code for image annotation. The IRMA code is currently available in English and German. It is planned to use the results of such automatic image annotation tasks for further, textual image retrieval tasks in the future. However, to simplify the task, only 57 simple class numbers are provided for ImageCLEF 2005. The meaning of each class and the number of images belonging to it is shown in Table 1.

Table 1: Annotation classes

Class	Description			Images	%
01	Radiography	Coronal	cranium, musculosceletal system	336	3.7%
02			cranium, facial cranium, musculosceletal system	32	0.4%
03			spine, cervical spine, musculosceletal system	215	2.4%
04			spine, thoracic spine, musculosceletal system	102	1.1%
05			spine, lumbar spine, musculosceletal system	225	2.5%
06			arm, hand, musculosceletal system	576	6.4%
07			arm, radio carpal joint, musculosceletal system	77	0.9%
08			arm, handforearm, musculosceletal system	48	0.5%
09			arm, elbow, musculosceletal system	69	0.8%
10			arm, upper arm, musculosceletal system	32	0.4%
11			arm, shoulder, musculosceletal system	108	1.2%
12			chest	2563	28.5%
13			chest, bones, musculosceletal system	93	1.0%
14			abdomen, gastrointestinal system	152	1.7%
15			abdomen, uropoietic system	15	0.2%
16			abdomen, upper abdomen, gastrointestinal system	23	0.3%
17			pelvis, musculosceletal system	217	2.4%
18			leg, foot, musculosceletal system	205	2.3%
19			leg, ankle joint, musculosceletal system	137	1.5%
20			leg, lower leg, musculosceletal system	31	0.3%
21			leg, knee, musculosceletal system	194	2.2%
22			leg, upper leg, musculosceletal system	48	0.5%
23			leg, hip, musculosceletal system	79	0.9%
24		Sagittal	cranium, facial cranium, musculosceletal system	17	0.2%
25			cranium, neuro cranium, musculosceletal system	284	3.2%
26			spine, cervical spine, musculosceletal system	170	1.9%

27			spine, thoracic spine, musculosceletal system	109	1.2%
28			spine, lumbar spine, musculosceletal system	228	2.5%
29			arm, hand, musculosceletal system	86	1.0%
30			arm, radio carpal joint, musculosceletal system	59	0.7%
31			arm, forearm, musculosceletal system	60	0.7%
32			arm, elbow, musculosceletal system	78	0.9%
33			arm, shoulder, musculosceletal system	62	0.7%
34			chest	880	9.8%
35			leg, foot, musculosceletal system	18	0.2%
36			leg, ankle joint, musculosceletal system	94	1.0%
37			leg, lower leg, musculosceletal system	22	0.2%
38			leg, knee, musculosceletal system	116	1.3%
39			leg, upper leg, musculosceletal system	38	0.4%
40			leg, hip, musculosceletal system	51	0.6%
41		Axial	mamma, right breast, reproductive system	65	0.7%
42			mamma, left breast, reproductive system	74	0.8%
43			leg, knee, musculosceletal system	98	1.1%
44		Other orientation	cranium, facial cranium, musculosceletal system	193	2.1%
45			cranium, neuro cranium, musculosceletal system	35	0.4%
46			spine, cervical spine, musculosceletal system	30	0.3%
47			arm, hand, musculosceletal system	147	1.6%
48			mamma, right breast, reproductive system	79	0.9%
49			mamma, left breast, reproductive system	78	0.9%
50			leg, foot, musculosceletal system	91	1.0%
51	Fluoroscopy	Coronal	thorax, hilum, respiratory system	9	0.1%
52			abdomen, upper abdomen, gastrointestinal system	9	0.1%
53			pelvis, cardiovascular system	15	0.2%
54			leg, lower leg, cardiovascular system	46	0.5%
55			leg, knee, cardiovascular system	10	0.1%
56			leg, upper leg, cardiovascular system	15	0.2%
57	Angiography	Coronal	pelvis, cardiovascular system	57	0.6%

The distribution of images is not homogeneous among all classes, with a clear deviation to class 12 (chest) with more than 28% of the training images. This may be cause for concern when building the classifiers and should be taken into account.

3 Description of experiments

This task is clearly aimed at image analysis research groups and the areas of expertise of the MIRACLE group don’t include image analysis research. However, as our group did have a strong expertise in automatic learning algorithms applied to different projects mainly in the fields of data, text and web mining, we decided to make the effort and participate in this task to promote and encourage multidisciplinary participation in all aspects of information retrieval, no matter if it is text or content based.

To isolate from the content-based retrieval part of the process, we resorted to GIFT (GNU Image Finding Tool) [1], a publicly available content-based image retrieval system which was developed under the GNU license and allows to perform query by example on images, using an image as the starting point for the search process. GIFT relies entirely on the image contents and thus it doesn’t require the collection to be annotated. It also provides a mechanism to improve query results by relevance feedback.

Our approach is based on the multidisciplinary combination of the usage of GIFT to perform content-based searches and the application of learning techniques over the retrieval results to build a classifier. Our system is divided in two parts: the content-based retrieval component (mainly GIFT) and the learning component, which makes calls to the retrieval component when necessary and uses the results to build the classifier. We think that this is a naive approach in the sense that we had to completely trust the results from the retrieval engine without no possibility or knowledge to change its behaviour. The only margin for improvement was on the learning component of the system, which in fact relied on the retrieval component.

We finally submitted two different runs to be evaluated by the task coordinators.

Retrieval Component

Unzipping the database with the 9,000 training images provided by the task coordinators results in a structure of 57 directories (Train01 to Train57), which allows to easily know the class of each image simply by parsing the path of the file. GIFT was then used to index the whole set of images, down-scaled to 32x32 pixels.

The retrieval component takes two parameters as inputs: a query image and a number of results. It internally makes calls to GIFT with the image as a query, gets the images that are more similar to the query image, and finally returns the given number of top results, each with the filename of the image and its relevance.

Although different search algorithms can be integrated as plug-ins in GIFT, only the provided separate normalisation algorithm has been used in our experiments.

Decision Table Classifier

This run selects the classification by using a decision table majority classifier [5]. First, the retrieval component performs an initial search of the query image and returns a list of the top N images which their relevancies. Then, a weighting function is applied to the relevance of each result. Finally, as each result is associated to a particular class (which can be easily obtained just parsing the filename), the confidence of each class is calculated as the sum of the weighted relevancies of all the results which correspond to that class. The output is the confidence of each of the 57 classes, assuming that the class with the highest relevance is considered to be the class of the image.

After several tests using 10-fold cross-validation with the training images, the best results were obtained when assuming N=20 (taking the top 20 results to compute the aggregated class confidence) and using a factor of 1/n (n being the number of result, from 1 to 20) as the weighting function.

Nearest-Neighbour Classifier

This run is based on the previous experiment and applies a K-Nearest-Neighbour classifier [5] to predict the output class. The classifier is trained for all the training examples (images), using 58 input variables: the vector of confidences for the 57 classes (float values), calculated as explained before, and the class which corresponds to the maximum value (string value). The output variable (the one to model in training) is the class of each training image. This variable will be predicted later for each test image. Weka [5],[12] was used to implement this classifier.

After several tests using 10-fold cross-validation with the training images, the best results were obtained when assuming K=8 (8 nearest examples) and enabling attribute normalization and no distance weighting. Although we were aware of the non homogeneous training examples among different classes, we didn’t take this fact into account due to lack of time to carry out the experiments.

4 Evaluation

Figure 1 compares the class distribution of training/test images (respectively, 9,000/1,000 images), analyzing the qrels file provided by the task coordinator after the submission deadline. The full data is shown in Table 4 (see appendix).

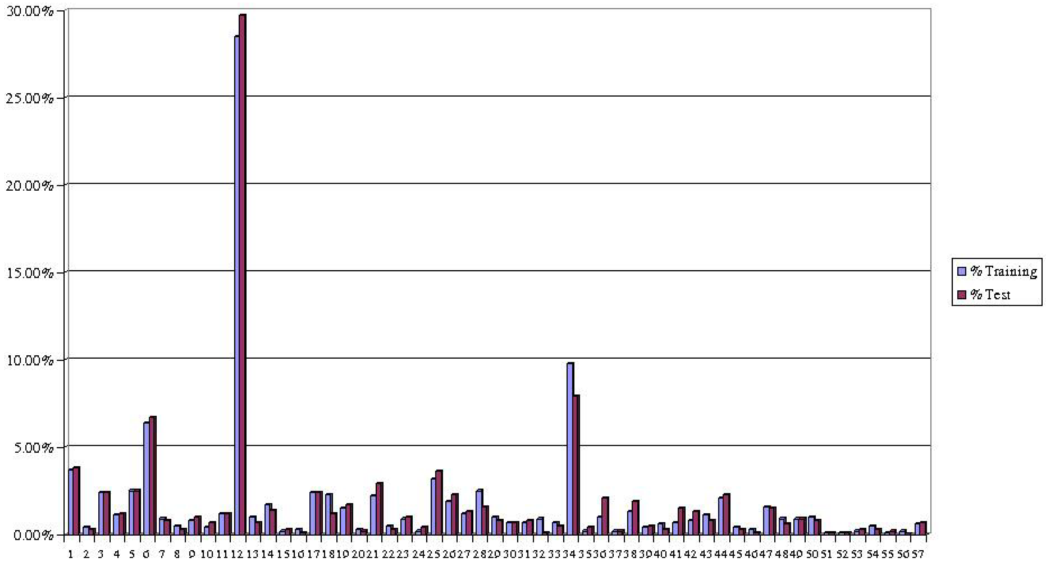

Figure 1: Comparison between training/test class distribution

As in the training set, some sampling bias can be observed also in the test set, and, furthermore, differences in the relative distribution between them. We think that this differences may affect the building of learning models and the also the evaluation of the different groups, and it has to be taken into account for next years.

The results of the classifiers are shown in Table 2, ranked by error rate (note that each 0.1% corresponds to 1 misclassification).

Table 2: Evaluation of classifiers

Run	Error rate
mira20relp57.txt (1)	21.4
mira20relp58IB8.txt (2)	22.3
Euclidean distance, 32x32 images, 1-Nearest-Neighbour (3)	36.8

(1) mira20relp57.txt is the Decision table classifier

(2) mira20relp58IB8.txt is the Nearest neighbour classifier

(3) According to the track organizers, for a 1-Nearest-Neighbour classifier comparing the images down-scaled to 32x32 pixels using Euclidean distance, the error rate is 36.8% (which means 368 images were misclassified).

As shown in Table 2, the best result was obtained with the decision table classifier. This error rate greatly improves the baseline of a 1-nearest-neighbour classifier.

The differences between our two runs haven’t still been analysed in detail at this moment, but a possible explanation for the performance loss with the nearest neighbour classifier may be imputed to model overtraining (when selecting the value of the parameters) or to the non homogeneous distribution of the training examples.

Figure 2 shows the distribution of error rate for each class for the decision table classifier. The full data is shown in Table 5 (see appendix). Note the lower error rate for the classes with many training examples (such as 12 and 34). This may also lead to think that the model is overfitting the training set.

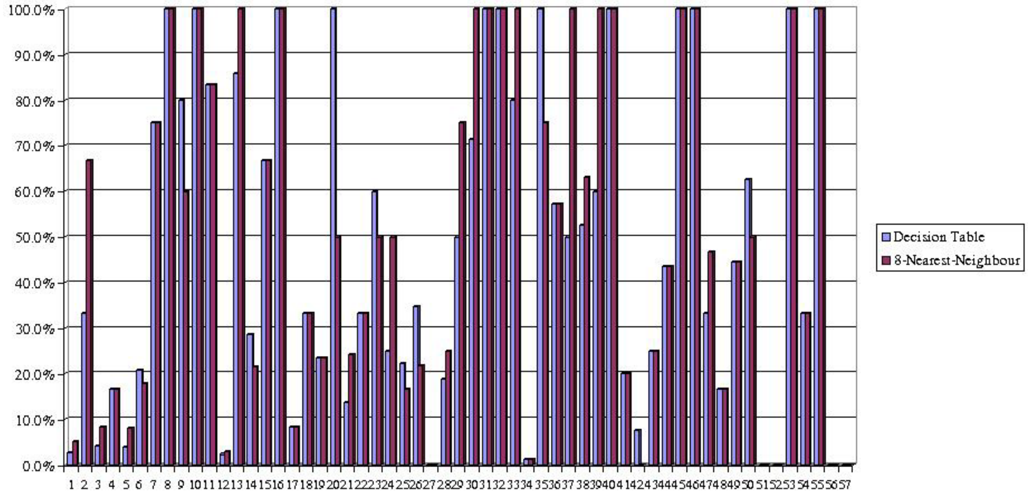

Figure 2: Error rate per class

Apart from MIRACLE, other 12 groups participated in this year's evaluation, with more than one submission. Next table shows the results for each group’s best submission.

Table 3: Comparison among groups

Group	Error rate	Difference
rwth-i6	12.6	
rwth-mi	13.3	0.7
ulg.ac.be	14.1	1.5
geneva-gift	20.6	8.0
infocomm	20.6	8.0
miracle	21.4	8.8
ntu	21.7	9.1
nctu-dblab	24.7	12.1
cea	36.9	24.3
mtholyoke	37.8	25.2
cindi	43.3	30.7
montreal	55.7	43.1

Our best submission misclassifies only 88 images more than the best submission, from RWTH Aachen Computer Science. These results are very satisfactory for us, considering the simple techniques which have been employed in our experiment and also that we are not a group with expertise in image processing as the others.

5 Conclusions and Future Work

Our naive approach to this task has proved to be valuable and the results are good enough to be comparable to more sophisticated techniques used by other groups. The combination of a “black-box” search using a publicly accessible content-based retrieval engine with a simple classification algorithm based on a decision table with weighted relevance aggregation has turned to provide similar results to other “more complex” algorithms such as nearest-neighbour or not much worse than boosting. This simplicity may be a good starting point for the implementation of a real system.

Regarding the classifier, we think that there is still space for improvement with a more careful training of the model, probably having a better selection of the training set, or introducing extra parameters to model the sampling biases. We will study if another combination of parameter values would have led to better results, controlling the model overtraining.

In addition, there are other techniques which we also want to test, such as decision trees, neural networks or Bayesian algorithms. Another possibility which couldn’t be tested due to lack of time is to apply clustering to filter results before the learning algorithm, which could discriminate noisy classes and thus increase precision.

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, Ana González-Ledesma, José Mª Guirao-Miras, Sara Lana-Serrano, Paloma Martínez-Fernández, Ángel Martínez-González, Antonio Moreno-Sandoval and César de Pablo Sánchez.

Appendix: Other tables

Table 4: Image distribution in training/test sets

Class	% Training [(1)]	% Test [(2)]	Difference
01	3.70%	3.8%	0.1
02	0.40%	0.3%	-0.1
03	2.40%	2.4%	0.0
04	1.10%	1.2%	0.1
05	2.50%	2.5%	0.0
06	6.40%	6.7%	0.3
07	0.90%	0.8%	-0.1
08	0.50%	0.3%	-0.2
09	0.80%	1.0%	0.2
10	0.40%	0.7%	0.3
11	1.20%	1.2%	0.0
12	28.50%	29.7%	1.2
13	1.00%	0.7%	-0.3
14	1.70%	1.4%	-0.3
15	0.20%	0.3%	0.1
16	0.30%	0.1%	-0.2
17	2.40%	2.4%	0.0
18	2.30%	1.2%	-1.1
19	1.50%	1.7%	0.2
20	0.30%	0.2%	-0.1
21	2.20%	2.9%	0.7
22	0.50%	0.3%	-0.2
23	0.90%	1.0%	0.1
24	0.20%	0.4%	0.2
25	3.20%	3.6%	0.4
26	1.90%	2.3%	0.4
27	1.20%	1.3%	0.1
28	2.50%	1.6%	-0.9
29	1.00%	0.8%	-0.2
30	0.70%	0.7%	0.0
31	0.70%	0.8%	0.1
32	0.90%	0.1%	-0.8
33	0.70%	0.5%	-0.2
34	9.80%	7.9%	-1.9
35	0.20%	0.4%	0.2
36	1.00%	2.1%	1.1
37	0.20%	0.2%	0.0
38	1.30%	1.9%	0.6
39	0.40%	0.5%	0.1
40	0.60%	0.3%	-0.3
41	0.70%	1.5%	0.8
42	0.80%	1.3%	0.5
43	1.10%	0.8%	-0.3
44	2.10%	2.3%	0.2
45	0.40%	0.3%	-0.1
46	0.30%	0.1%	-0.2

47	1.60%	1.5%	-0.1
48	0.90%	0.6%	-0.3
49	0.90%	0.9%	0.0
50	1.00%	0.8%	-0.2
51	0.10%	0.1%	0.0
52	0.10%	0.1%	0.0
53	0.20%	0.3%	0.1
54	0.50%	0.3%	-0.2
55	0.10%	0.2%	0.1
56	0.20%	0.0%	-0.2
57	0.60%	0.7%	0.1

Table 5: Confusion matrix for both classifiers

Class	Right	Wrong	Error Rate	Related classes
01	37	1	2.6%	2
02	2	1	33.3%	2
03	23	1	4.2%	2
04	10	2	16.7%	3
05	24	1	4.0%	2
06	53	14	20.9%	9
07	2	6	75.0%	6
08	0	3	100.0%	2
09	2	8	80.0%	5
10	0	7	100.0%	5
11	2	10	83.3%	10
12	290	7	2.4%	7
13	1	6	85.7%	5
14	10	4	28.6%	4
15	1	2	66.7%	3
16	0	1	100.0%	1
17	22	2	8.3%	3
18	8	4	33.3%	3
19	13	4	23.5%	5
20	0	2	100.0%	2
21	25	4	13.8%	4
22	2	1	33.3%	2
23	4	6	60.0%	7
24	3	1	25.0%	2
25	28	8	22.2%	6
26	15	8	34.8%	7
27	13	0	0.0%	1
28	13	3	18.8%	4
29	4	4	50.0%	4
30	2	5	71.4%	5
31	0	8	100.0%	6
32	0	1	100.0%	1
33	1	4	80.0%	5
34	78	1	1.3%	2
35	0	4	100.0%	4
36	9	12	57.1%	9
37	1	1	50.0%	2
38	9	10	52.6%	7
39	2	3	60.0%	4
40	0	3	100.0%	2
41	12	3	20.0%	2
42	12	1	7.7%	2
43	6	2	25.0%	3
44	13	10	43.5%	9
45	0	3	100.0%	2
46	0	1	100.0%	1

47	10	5	33.3%	3
48	5	1	16.7%	2
49	5	4	44.4%	2
50	3	5	62.5%	5
51	1	0	0.0%	1
52	1	0	0.0%	1
53	0	3	100.0%	2
54	2	1	33.3%	2
55	0	2	100.0%	2
56	0	0	0.0%	0
57	7	0	0.0%	1

The last column shows the number of different classes related to the test class, which can be used to study the noise level for the class or its discrimination ability.

References

[1] GIFT: The GNU Image-Finding Tool. On line http://www.gnu.org/software/gift/ [Visited 18/07/2005]

[2] 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).

[3] 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.

[4] Goodrum, A.A.: Image Information Retrieval: An Overview of Current Research. Informing Science, Vol 3(2):63-66 (2000).

[5] Ian H. Witten and Eibe Frank: Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[6] IRMA project: Image Retrieval in Medical Applications. On line http://www.irma-project.org/ [Visited 18/07/2005]

[7] 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).

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Weka: Data Mining Software in Java. On line http://www.cs.waikato.ac.nz/ml/weka/ [Visited 18/07/2005]