The University of Jaén at ImageCLEF 2005: Adhoc and Medical Tasks

M.T. Martín-Valdivia[1], M.A. García-Cumbreras[1], M.C. Díaz-Galiano[1], L.A. Ureña-López[1], and A. Montejo-Raez[1]

University of Jaén. Departamento de Informática
Grupo Sistemas Inteligentes de Acceso a la Información
Campus Las Lagunillas, Ed. A3, e-23071, Jaén, Spain
{maite, magc, mcdiaz, laurena, amontejo}@ujaen.es

Abstract. In this paper, we describe our first participation in the ImageCLEF campaign. The SINAI research group participated in both the ad hoc task and the medical task. For the first task, we have used several translation schemes as well as experiments with and without Pseudo Relevance Feedback (PRF). A voting-based system has been developed, for the ad hoc task, joining three different systems of participant Universities. For the medical task, we have also submitted runs with and without PRF, and experiments using only textual query and using textual mixing with visual query.

1 Introduction

This is the first participation for the SINAI research group at the ImageCLEF competition. We have accomplished the ad hoc task and the medical task [1].

As a cross language retrieval task, a multilingual image retrieval based on query translation can achieve high performance, more than a monolingual retrieval. The ad hoc task involves to retrieve relevant images using the text associated with each image query.

The goal of the medical task is to retrieve relevant images based on an image query. This year, a short text is associated with each image query. We first compare the results obtained using only textual query versus results obtained combining textual and visual information. We have accomplished several runs with and without Pseudo Relevance Feedback (PRF). Finally, we have used different methods to merge visual and text results.

The next section describes the ad hoc experiments. In Section 3, we explain the experiments for the medical task. Finally, conclusions and proposals for work are presented in Section 4.

2 The Ad Hoc Task

The goal of the ad hoc task is, given a multilingual query, to find as many relevant images as possible from an image collection.

C. Peters et al. (Eds.): CLEF 2005, LNCS 4022, pp. 612–621, 2006.

The proposal of the ad hoc task is to compare results with and without PRF, with or without query expansion, using different methods of query translation or using different retrieval models and weighting functions.

2.1 Experiment Description

In our experiments we have used nine languages: English, Dutch, Italian, Spanish, French, German, Danish, Swedish, and Russian. The dataset is the same used in 2004: St Andrews. The St Andrews dataset consists of 28,133 photographs from the St Andrews University Library photographic collection which holds one of the largest and most important collections of historic photography in Scotland. The collection numbers in excess of 300,000 images, 10% of which have been digitized and used for the ImageCLEF ad hoc retrieval task. All images have an accompanying textual description consisting of 8 distinct fields. These fields can be used individually or collectively to facilitate image retrieval. The collections have been preprocessed, using stopwords and the Porters stemmer.

The collection has been indexed using LEMUR IR system[1], it is a toolkit that supports indexing of large-scale text databases, the construction of simple language models for documents, queries, or subcollections, and the implementation of retrieval systems based on language models as well as a variety of other retrieval models.

We have used online Machine Translator for each language pair, using English as pivot language. One parameter for each experiment is the weighting function, such as Okapi or TFIDF. Another is the use or not of PRF.

2.2 Results and Discussion

Tables 1, 2, 3, 4, 5, 6, 7, 8, and 9 show a summary of experiments submitted and results obtained for the seven languages used.

The results obtained show that in general the application of query expansion improves the results. Only one Italian experiment without query expansion gets a better result. In the case of the use of only title or title + narrative, the results are not conclusive, but the use of only title seems to produce better results.

Table 1. Summary of results for the ad hoc task (Dutch)

Experiment	Initial Query	Expansion	MAP	%MONO	Rank
SinaiDuTitleFBSystran	title	with	0.3397	66.5%	2/15
SinaiDuTitleNoFBSystran	title	without	0.2727	53.4%	9/15

2.3 Joint Participation

For the ad hoc task we have also made a joint participation within the R2D2 project framework. We have integrated our system and the ones belonging to the UNED group from Madrid and the system from the University of Alicante (UA).

[1] http://www.lemurproject.org/

Table 2. Summary of results for the ad hoc task (English)

Experiment	Initial Query	Expansion	MAP	%MONO	Rank
SinaiEnTitleNarrFB	title + narr	with	0.3727	n/a	31/70
SinaiEnTitleNoFB	title	without	0.3207	n/a	44/70
SinaiEnTitleFB	title	with	0.3168	n/a	45/70
SinaiEnTitleNarrNoFB	title + narr	without	0.3135	n/a	46/70

Table 3. Summary of results for the ad hoc task (French)

Experiment	Initial Query	Expansion	MAP	%MONO	Rank
SinaiFrTitleNarrFBSystran	title + narr	with	0.2864	56.1%	1/17
SinaiFrTitleNarrNoFBSystran	title + narr	without	0.2227	43.6%	12/17
SinaiFrTitleFBSystran	title	with	0.2163	42.3%	13/17
SinaiFrTitleNoFBSystran	title	without	0.2158	42.2%	14/17

Table 4. Summary of results for the ad hoc task (German)

Experiment	Initial Query	Expansion	MAP	%MONO	Rank
SinaiGerTitleFBSystran	title	with	0.3004	58.8%	4/29
SinaiGerTitleFBPrompt	title	with	0.2931	57.4%	5/29
SinaiGerTitleNoFBPrompt	title	without	0.2917	57.1%	6/29
SinaiGerTitleNarrFBSystran	title + narr	with	0.2847	55.7%	7/29
SinaiGerTitleNarrFBPrompt	title + narr	with	0.2747	53.8%	10/29
SinaiGerTitleNoFBSystran	title	without	0.2720	53.2%	13/29
SinaiGerTitleFBWordlingo	title	with	0.2491	48.8%	16/29
SinaiGerTitleNarrNoFBSystran	title + narr	without	0.2418	47.3%	17/29
SinaiGerTitleNarrNoFBPrompt	title + narr	without	0.2399	47.0%	18/29
SinaiGerTitleNoFBWordlingo	title	without	0.2217	43.4%	19/29
SinaiGerTitleNarrFBWordlingo	title + narr	with	0.1908	37.4%	21/29
SinaiGerTitleNarrNoFBSWordlingo	title + narr	without	0.1860	36.4%	22/29

Table 5. Summary of results for the ad hoc task (Italian)

Experiment	Initial Query	Expansion	MAP	%MONO	Rank
SinaiItTitleNoFBSystran	title	without	0.1805	35.3%	12/19
SinaiItTitleFBSystran	title	with	0.1672	32.7%	13/19
SinaiItTitleNarrNoFBSystran	title + narr	without	0.1585	31.0%	14/19
SinaiItTitleNoFBWordlingo	title	without	0.1511	29.6%	15/19
SinaiItTitleNarrFBSystran	title + narr	with	0.1397	27.3%	16/19
SinaiItTitleFBWordlingo	title	with	0.1386	27.1%	18/19

Table 6. Summary of results for the ad hoc task (Russian)

Experiment	Initial Query	Expansion	MAP	%MONO	Rank
SinaiRuTitleFBSystran	title	with	0.2229	43.6%	11/15
SinaiRuTitleNoFBSystran	title	without	0.2096	41.0%	12/15

Table 7. Summary of results for the ad hoc task (Spanish European)

Experiment	Initial Query	Expansion	MAP	%MONO	Rank
SinaiSpEurTitleFBPrompt	title	with	0.2416	47.3%	5/33
SinaiSpEurTitleFBEpals	title	with	0.2292	44.9%	7/33
SinaiSpEurTitleNoFBPrompt	title	without	0.2260	44.2%	8/33
SinaiSpEurTitleNarrFBEpals	title + narr	with	0.2135	41.8%	11/33
SinaiSpEurTitleNoFBEpals	title	without	0.2074	40.6%	16/33
SinaiSpEurTitleNarrFBSystran	title + narr	with	0.2052	40.2%	20/33
SinaiSpEurTitleNoFBSystran	title	without	0.1998	39.1%	21/33
SinaiSpEurTitleNoFBWordlingo	title	without	0.1998	39.1%	22/33
SinaiSpEurTitleFBSystran	title	with	0.1965	38.5%	23/33
SinaiSpEurTitleFBWordlingo	title	with	0.1965	38.5%	24/33
SinaiSpEurTitleNarrNoFBEpals	title + narr	without	0.1903	37.3%	25/33
SinaiSpEurTitleNarrNoFBPrompt	title + narr	without	0.1865	36.5%	27/33
SinaiSpEurTitleNarrNoFBSystran	title + narr	without	0.1712	33.5%	28/33
SinaiSpEurTitleNarrFBSystran	title + narr	with	0.1605	31.4%	29/33
SinaiSpEurTitleNarrNoFBSWordlingo	title + narr	without	0.1343	26.3%	31/33
SinaiSpEurTitleNarrFBWordlingo	title + narr	with	0.1182	23.1%	32/33

Table 8. Summary of results for the ad hoc task (Spanish Latinamerican)

Experiment	Initial Query	Expansion	MAP	%MONO	Rank
SinaiSpLatTitleFBPrompt	title	with	0.2967	58.1%	8/31
SinaiSpLatTitleNoFBPrompt	title	without	0.2963	58.0%	9/31
SinaiSpLatTitleNoFBEpals	title	without	0.2842	55.6%	11/31
SinaiSpLatTitleNoFBSystran	title	without	0.2834	55.5%	12/31
SinaiSpLatTitleNoFBWordlingo	title	without	0.2834	55.5%	13/31
SinaiSpLatTitleFBSystran	title	with	0.2792	54.7%	14/31
SinaiSpLatTitleFBWordlingo	title	with	0.2792	54.7%	15/31
SinaiSpLatTitleFBEpals	title	with	0.2606	51.0%	16/31
SinaiSpLatTitleNarrNoFBSystran	title + narr	without	0.2316	45.3%	19/31
SinaiSpLatTitleNarrFBPrompt	title + narr	with	0.2259	44.2%	20/31
SinaiSpLatTitleNarrFBSystran	title + narr	with	0.2026	39.7%	21/31
SinaiSpLatTitleNarrFBEpals	title + narr	with	0.2001	39.2%	22/31
SinaiSpLatTitleNarrNoFBPrompt	title + narr	without	0.1992	39.0%	23/31
SinaiSpLatTitleNarrNoFBEpals	title + narr	without	0.1900	37.2%	24/31
SinaiSpLatTitleNarrNoFBSWordlingo	title + narr	without	0.1769	34.6%	25/31
SinaiSpLatTitleNarrFBWordlingo	title + narr	with	0.1459	28.6%	27/31

Table 9. Summary of results for the ad hoc task (Swedish)

Experiment	Initial Query	Expansion	MAP	%MONO	Rank
SinaiSweTitleNoFBSystran	title	without	0.2074	40.6%	2/7
SinaiSweTitleFBSystran	title	with	0.2012	39.4%	3/7

We have developed a voting system among them. For English, Dutch, French, German, Italian, Russian and Spanish we have done a combination between UA and SINAI. UA, UNED and SINAI systems have been only combined for Spanish. The parameters selected are the use of feedback and the use of query titles and automatic translation. The voting was developed using the weights of each document in each retrieved list.

The results are shown in the Table 10. The ranks are shown in brackets.

Table 10. Summary of results for the voting-based collaborative system

Language	SINAI	UA	UNED	SINAI-UA	SINAI-UA-UNED
Dutch	0.3397(2/15)	0.2765(8/15)	-	0.3435(1/15)	-
English	0.3727(30/70)	0.3966(14/70)	-	0.4080(7/70)	-
French	0.2864(1/17)	0.2621(6/17)	-	0.2630(5/17)	-
German	0.3004(4/29)	0.2854(7/29)	-	0.3375(1/29)	-
Italian	0.1805(11/19)	0.2230(4/19)	-	0.2289(2/19)	-
Russian	0.2229(11/15)	0.2683(3/15)	-	0.2665(5/15)	-
Spanish (eur)	0.2416(5/33)	0.2105(12/33)	0.3175(1/33)	0.2668(4/33)	0.3020(2/33)
Spanish (lat)	0.2967(8/31)	0.3179(2/31)	0.2585(17/31)	0.3447(1/31)	0.3054(4/31)

As we can see the voting system improves the results for the Dutch, English, German, Italian and Spanish-latinoamerican languages.

3 The Medical Task

The main goal of medical task is to improve the retrieval of medical images from heterogeneous and multilingual document collections containing images as well as text. This year, queries have been formulated with example images and a short textual description explaining the research goal. For the medical task, we have used the list of retrieved images by GIFT[2] [2] which was supplied by the organizers of this track. Also, we used the text of topics for each query. For this reason, our efforts concentrated on manipulating the text descriptions associated with these images and in mixing the partial results lists. Thus, our experiments only use the list provided by the GIFT system in order to expand textual queries. Textual descriptions of the medical cases have been used to try to improve retrieval results.

3.1 Textual Retrieval System

In order to generate the textual collection we have used the images and their annotations.

The entire collection consists of 4 datasets (CASImage, Pathopic, Peir and MIR) containing about 50,000 images. Each subcollection is organized into cases that represent a group of related images and annotations. Each case consists in

[2] `http://www.gnu.org/software/gift/`

a group of images and an optional annotation. The collection annotations are in XML format. The majority of the annotations are in English but a significant number is also in French (CASImage) and German (Pathopic), with a few cases that do not contain any annotation at all. The quality of the texts is variable between collections and even within the same collection.

We generated a textual document per image, where the identifier number of document is the name of the image and the text of document is the XML annotation associated with this image. The XML tags and unnecessary fields such as *LANGUAGE* were removed. If there were several images of the same case, the text was copied several times.

We have used English language for the document collection as well for the queries. Thus, French annotations in CASImage collection were translated to English and then were incorporated with the collection. Pathopic collection has annotation in both English and German language. We only used English annotations in order to generate the Pathopic documents and German annotations were discarded.

Finally, we have added the text associated with each query topic as documents. In this case, if a query topic includes several images, the text was also copied several times.

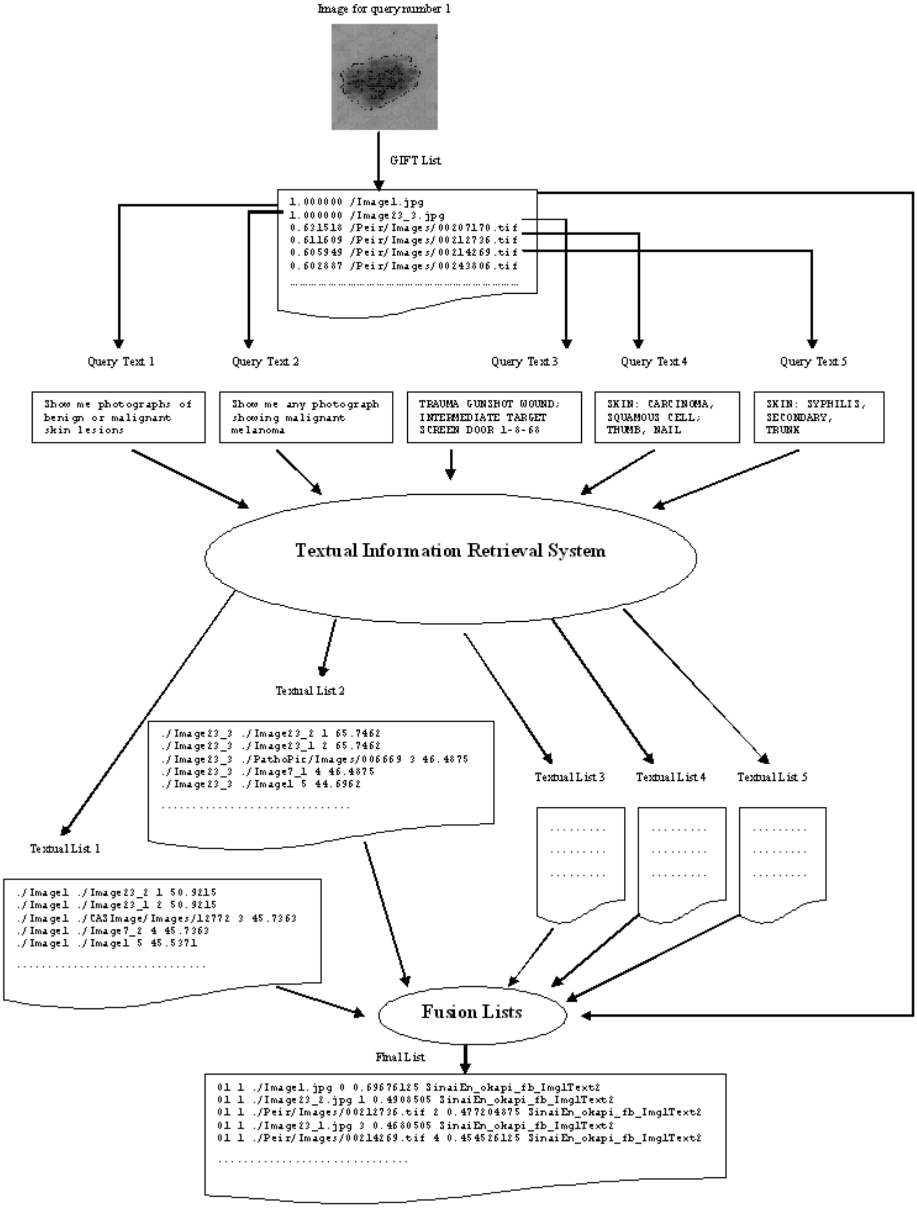
Once the document collection was generated, experiments were conducted with the LEMUR retrieval information system. We have used the 3 different weighting schemes available: TFIDF, Okapi and Kl-divergence.

3.2 Experiment Description

Our main goal is to investigate the effectiveness of combining text and image for retrieval. For this, we compare the obtained results when we only use the text associated with the query topic and the results when we merge visual and textual information.

We have accomplished a first experiment that we have used as baseline case. This experiment simply consists of taking the text associated with each query as a new textual query. Then, each textual query is submitted to the LEMUR system. The resulting list is directly the baseline run. This result list from LEMUR system contains the most similar cases with respect to the text and a weighting (the relevance). The weighting was normalized based on the highest weighting in the list to get values between 0 and 1.

The remaining experiments start from the ranked lists provided by the GIFT. The organization provides a GIFT list of relevant images for each query. For each list/query we have used an automatic textual query expansion of the first five images from the GIFT lists. We have taken the text associated with each image in order to generate a new textual query. Then, each textual query is submitted to the LEMUR system and we obtain five new ranked lists. Again, the resulting lists were normalized to 1. Thus, for each original query we have six partial lists. The last step consists of merging these partial result lists using some strategy in order to obtain one final list with relevant images ranking by relevance. Figure 1 describes the process.



```
1.000000 /Image1.jpg
1.000000 /Image23_3.jpg
0.631518 /Peir/Images/00207170.tif
0.611609 /Peir/Images/00212736.tif
0.605949 /Peir/Images/00214269.tif
0.602887 /Peir/Images/00243806.tif
..............................................
```

Show me photographs of benign or malignant skin lesions

Show me any photograph showing malignant melanoma

TRAUMA GUNSHOT WOUND; INTERMEDIATE TARGET SCREEN DOOR 1-8-68

SKIN: CARCINOMA, SQUAMOUS CELL; THUMB, NAIL

SKIN: SYPHILIS, SECONDARY, TRUNK

```
./Image23_3 ./Image23_2 1 65.7462
./Image23_3 ./Image23_1 2 65.7462
./Image23_3 ./PathoPic/Images/006669 3 46.4875
./Image23_3 ./Image7_1 4 46.4875
./Image23_3 ./Image1 5 44.6962
..............................
```

```
./Image1 ./Image23_2 1 50.9215
./Image1 ./Image23_1 2 50.9215
./Image1 ./CASImage/Images/12772 3 45.7363
./Image1 ./Image7_2 4 45.7363
./Image1 ./Image1 5 45.5371
..............................
```

```
.........
.........
.........
```

```
.........
.........
.........
```

```
.........
.........
.........
```

```
01 1 ./Image1.jpg 0 0.69676125 SinaiEn_okapi_fb_Img1Text2
01 1 ./Image23_2.jpg 1 0.4908505 SinaiEn_okapi_fb_Img1Text2
01 1 ./Peir/Images/00212736.tif 2 0.477204875 SinaiEn_okapi_fb_Img1Text2
01 1 ./Image23_1.jpg 3 0.4680505 SinaiEn_okapi_fb_Img1Text2
01 1 ./Peir/Images/00214269.tif 4 0.454526125 SinaiEn_okapi_fb_Img1Text2
..............................
```

Fig. 1. The merging process of result lists

The merging of the visual and textual results was done in various ways:

1. **ImgText4:** The final list includes the images present in at least 4 partial lists independently of these lists are visual or textual. In order to calculate the final image relevance simply we sum the partial relevance and divide by the maximum number of lists where the images are present.
2. **ImgText3:** This experiment is the same as ImgText4 but the image must be in at least 3 lists.
3. **ImgText2:** This experiment is the same as ImgText4 but the image must be in at least 2 lists.
4. **Img1Tex4:** The final list includes the images present in at least 4 partial lists, but the image is necessary to be in the GIFT list (i.e., the image must be in the GIFT list and in at least other 3 textual lists). In order to calculate the final image relevance we simply sum the partial relevance and divide by the maximum number of lists where the images are present.
5. **Img1Text3:** This experiment is the same as Img1Text4, but the image must be in at least 3 lists (the GIFT list and at least 2 textual lists).
6. **Img1Text2:** This experiment is the same as Img1Text4, but the image must be in at least 2 lists (the GIFT list and at least 1 textual list).

These 6 experiments and the baseline experiment (that only uses textual information of the query) have been accomplished with and without PRF for each weighting schemes (TFIDF, Okapi and Kl-divergence). In summary, we have submitted 42 runs: 7 (different experiments)*2 (PRF and no PRF) * 3 (weighting schemes).

3.3 Results and Discussion

Tables 11 and 12 show the official results for medical task (text only and mixed retrieval). The total runs submitted for text were only 14 and for mixed retrieval 86. Best results were obtained when using Okapi without PRF for *text only* runs (experiment SinaiEn_okapi_nofb_Topics.imageclef2005) and using Kl-divergence with PRF and ImgText2 experiment for *mixed retrieval* runs (experiment SinaiEn_kl_fb_ImgText2.imageclef2005).

Table 11. Performance of official runs in Medical Image Retrieval (text only)

Experiment	Precision	Rank
IPALI2R_TIan (best result)	0.2084	1
SinaiEn_okapi_nofb_Topics.imageclef2005	0.091	5
SinaiEn_okapi_fb_Topics.imageclef2005	0.0862	6
SinaiEn_kl_fb_Topics.imageclef2005	0.079	7
SinaiEn_kl_nofb_Topics.imageclef2005	0.0719	8
SinaiEn_tfidf_fb_Topics.imageclef2005	0.0405	10
SinaiEn_tfidf_nofb_Topics.imageclef2005	0.0394	12

Table 12. Performance of official runs in Medical Image Retrieval (mixed text+visual)

Experiment	Precision	Rank
IPALI2R_Tn (best result)	0.2084	1
SinaiEn_kl_fb_ImgText2.imageclef2005	0.1033	24
SinaiEn_kl_fb_Img1Text2.imageclef2005	0.1002	28
SinaiEn_okapi_fb_Img1Text2.imageclef2005	0.0992	31
SinaiEn_okapi_nofb_Img1Text2.imageclef2005	0.0955	33
SinaiEn_kl_nofb_ImgText2.imageclef2005	0.0947	34
SinaiEn_okapi_nofb_ImgText2.imageclef2005	0.0931	36
SinaiEn_okapi_fb_ImgText2.imageclef2005	0.0905	39
SinaiEn_kl_fb_ImgText3.imageclef2005	0.0891	41
SinaiEn_kl_nofb_Img1Text2.imageclef2005	0.0884	42
SinaiEn_okapi_nofb_ImgText3.imageclef2005	0.0867	43
SinaiEn_kl_fb_Img1Text3.imageclef2005	0.0845	44
SinaiEn_okapi_fb_ImgText3.imageclef2005	0.0803	47
SinaiEn_kl_nofb_ImgText3.imageclef2005	0.0781	48
SinaiEn_okapi_nofb_Img1Text3.imageclef2005	0.0779	49
SinaiEn_okapi_fb_Img1Text3.imageclef2005	0.0761	50
SinaiEn_kl_nofb_Img1Text3.imageclef2005	0.0726	52
SinaiEn_okapi_nofb_ImgText4.imageclef2005	0.0685	53
SinaiEn_tfidf_fb_Img1Text2.imageclef2005	0.0678	54
SinaiEn_kl_nofb_ImgText4.imageclef2005	0.0653	57
SinaiEn_kl_nofb_Img1Text4.imageclef2005	0.0629	59
SinaiEn_kl_fb_ImgText4.imageclef2005	0.062	60
SinaiEn_kl_fb_Img1Text4.imageclef2005	0.0602	61
SinaiEn_okapi_nofb_Img1Text4.imageclef2005	0.0596	62
SinaiEn_tfidf_nofb_Img1Text2.imageclef2005	0.0582	63
SinaiEn_okapi_fb_Img1Text4.imageclef2005	0.055	64
SinaiEn_okapi_fb_ImgText4.imageclef2005	0.0547	65
SinaiEn_tfidf_fb_ImgText2.imageclef2005	0.0481	69
SinaiEn_tfidf_fb_Img1Text3.imageclef2005	0.0474	70
SinaiEn_tfidf_fb_ImgText3.imageclef2005	0.0713	76
SinaiEn_tfidf_nofb_Img1Text3.imageclef2005	0.0412	77
SinaiEn_tfidf_nofb_ImgText2.imageclef2005	0.0395	79
SinaiEn_tfidf_fb_ImgText4.imageclef2005	0.0386	80
SinaiEn_tfidf_fb_Img1Text4.imageclef2005	0.0372	82
SinaiEn_tfidf_nofb_ImgText3.imageclef2005	0.0362	83
SinaiEn_tfidf_nofb_Img1Text4.imageclef2005	0.0339	84
SinaiEn_tfidf_nofb_ImgText4.imageclef2005	0.0336	85

There are no significant differences between results obtained with Okapi and Kl-divergence schemes. However, the worst results were obtained with the TFIDF scheme.

On the other hand, the use of only two lists is better than mixing three or four lists of partial results. A substantial difference in the inclusion or not of the images in the GIFT list (Img1Text$X$ experiments) is not appraised, either.

4 Conclusion and Further Works

In this paper, we have presented the experiment carried out in our first participation in the ImageCLEF campaign. We have only tried to verify if the use of textual information increases the effectiveness of the systems. Evaluation results show that the use of textual information significantly improves the retrieval.

The incorporation of some natural language processing techniques such as word sense disambiguation (WSD) or named entity recognition (NER) will focus our future work. We also plan to use some machine learning algorithms in order to improve the lists merging process. Thus, we should do a comparative study for different fusion methods using basic algorithms (such as Round-Robin or Raw Scoring) and machine learning algorithms (such as logistic regression, neural networks and support vector machines).

Acknowledgements

This work has been partially supported by a grant from the Spanish Government, project R2D2 (TIC2003-07158-C04-04).

References

1. Clough P., Henning Mller, Thomas Deselaers, Michael Grubinger, Thomas M. Lehmann, Jeffery Jensen, William Hersh, The CLEF 2005 Cross-Language Image Retrieval Track, Proceedings of the Cross Language Evaluation Forum 2005, Springer Lecture Notes in Computer science, 2006 - to appear.
2. Müller, H., Geissbhler, A., Ruch., P.: Report on the CLEF experiment: Combining image and multi-lingual search medical image retrieval. In Proceedings of the Cross Language Evaluation Forum (CLEF 2004), 2004