SINAI at ImageCLEF 2006

M.C. Díaz-Galiano, M.A. García-Cumbreras, M.T. Martín-Valdivia,
A. Montejo-Raez, L.A. Ureña-López
University of Jaén. Departamento de Informática
Grupo Sistemas Inteligentes de Acceso a la Información
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{mcdiaz,magc,maite,amontejo,laurena}@ujaen.es

Abstract

This paper describes SINAI team participation in the ImageCLEF campaign. The SINAI research group participated in both the ad hoc task and the medical task. The experiments accomplished in both tasks result from very different approaches.

For the adhoc task the main IR system used is the same as that of the 2005 ImageCLEF adhoc task. The improvement of the adhoc system is a new Machine Translation system that works with several translators and implements several heuristics. We have participated in the English monolingual task and in six bilingual tasks for the languages: Dutch, French, German, Italian, Portuguese and Spanish. The results obtained shown that the English monolingual results are good (0,2234 is our best result) and there is a loss of precision with the bilingual runs and some languages like German or Spanish works better than others, because of the translations.

For the medical task, this year we carried out new and very different experiments to imageCLEFmed2005 ones. First of all, we have processed the set of collections using Information Gain (IG) to determine which are the best tags that should be considered in the indexing process. These tags are those supposed to provide the most relevant and non-redundant information, and have been selected automatically according to our information-based strategy along with the data and relevance assessments from last year.

This year, our goal was to analyze how tag selection may contribute to the quality of final results. In order to select reduced set of tags we have computed IG. 11 different collections were generated according to the percentage of tags with highest IG value. Finally, only results related to experiments with selections over the 20%, 30% and 40% of available tags were submitted, since they reported best performance on 2005 data.

Experiments using only textual query and using textual mixing with visual query have been submitted. For visual query we have used the GIFT lists provide by the organization. Surprisingly, the system performs better on the text retrieval alone than mixed textual and visual retrieval.

On the other hand, we try show that information filtering through tag selection using information gain improves retrieval results without the need of a manual selection, but the obtained results are no conclusive. Unfortunately, the results obtained are not as successful as desired. Due to a computing processing mistake all our mixed runs obtain the same results than the visual GIFT baseline (0.0467). At the moment of writing of this paper we are modifying our system in order to solve this problem.

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

Keywords

Visual and text retrieval, Information Gain, Indexing, Machine Translators

1 Introduction

This is the second participation of the SINAI research group at the ImageCLEF campaign. We have participated in the ad hoc task and the medical task.

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

The goal of the medical task is to retrieve relevant images based on an image query [1]. For this, organizers supply a multilingual and visual collection and a set of queries (images and a short text in English, French and German are associated). We first preprocess the collection using Information Gain (IG). This year, our main goal is to compare the effect of select different tags from the collection using this measure. We have attempted to choose those tags, providing the best information in order to improve the result obtained. We have generated several collections with different number of tags depending on their IG. Finally, we have only submitted runs on 3 different collections (at 20%,30% and 40%) because they reported the best results for the ImageCLEFmed2005 data. For each collection, we first compare the results obtained using only textual query against results obtained combining textual and visual information. Finally, we have used different methods to merge visual and textual results.

Next section describes the ad hoc experiments. In Section 3, we explain the experiments for the medical task. Finally, conclusions and future 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.

The proposal of the ad hoc task is to compare results with and without pseudo-relevant feedback, with or without query expansion, using different methods of query translation or using different retrieval models and weighting functions [2].

2.1 Experiments Description

In our experiments we have used seven languages: Dutch, English, French, German, Italian, Portuguese and Spanish.

Because in 2005 the results were quite good, this year we have used the same IR system and the same strategies, but introducing a new translation module. This module combines some Machine Translators and implements some heuristics.

The Machine Translators used have been (in brackets the translator by default for each language):

- Epals (German and Portuguese)
- Prompt (Spanish)
- Reverso (French)
- Systran (Dutch and Italian)

Some heuristics are, for instance, the use of the translation made by the translator by default, a combination with the translations of every translator, or a combination of the words with a higher punctuation (two points if it appears in the default translation and one point if it appears in all of the other translations).

Experiment	Initial Query	Expansion	Weight	MAP	Rank
sinaiEnEnFbOkapiExp1	title + narr	with	Okapi	0.2234	9/49
sinaiEnEnFbOkapiExp2	title + narr	without	Okapi	0.0845	38/49
sinaiEnEnFbOkapiExp3	title + narr	with	Tfidf	0.0846	37/49
sinaiEnEnFbOkapiExp4	title + narr	without	Tfidf	0.0823	39/49

Table 1: Summary of results for the English monolingual adhoc runs

Experiment	Initial Query	Expansion	Weight	MAP	Rank
sinaiDeEnFbOkapiExp1	title + narr	with	Okapi	0.1602	4/8
sinaiDeEnFbOkapiExp2	title + narr	without	Okapi	0.1359	7/8
sinaiDeEnFbOkapiExp3	title + narr	with	Tfidf	0.1489	5/8
sinaiDeEnFbOkapiExp4	title + narr	without	Tfidf	0.1369	6/8

Table 2: Summary of results for the German-English bilingual adhoc runs

The dataset is a new collection: IAPR. The IAPR TC-12 image collection consists of 20,000 images taken from locations around the world and comprising a varying cross-section of still natural images. It includes pictures of a range of sports, actions, photographs of people, animals, cities, landscapes and many other aspects of contemporary life.

The collections have been preprocessed, using stopwords and the Porter's stemmer.

The collection dataset has been indexed using LEMUR IR system. 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. The toolkit is being developed as part of the Lemur Project, a collaboration between the Computer Science Department at the University of Massachusetts and the School of Computer Science at Carnegie Mellon University.

One parameter for each experiment is the weighting function, such as Okapi or TFIDF. Another is the use or not of PRF (*pseudo-relevance feedback*).

2.2 Results and Discussion

As parameters all the results are obtained using the title and narrative text, when possible. In the English monolingual task and in the German-English bilingual task we have combined the use or not of pseudo-relevance feedback and the weighting function (Okapi or Tfidf).

In table 1, we can see the English monolingual results. The results obtained show that the pseudo-relevance feedback is too important when Okapi is used as weighing function. The results with Tfidf and with Okapi without PRF are very poor.

Table 2 show a summary of experiments submitted and results obtained for the German-English bilingual runs. In this case we have combine the same parameters than in the monolingual task.

The results obtained show that there is a loss of MAP between the best monolingual experiment and this bilingual, around a 28%. Even though, the other results in the English monolingual task are quite worse compared to the German bilingual ones.

Finally, table 3 show a summary of experiments submitted and results obtained for the other five bilingual runs.

The results obtained show that in general there is a loss of precision compared to the English monolingual results. The Spanish result is around a 17% worse. The other languages decrease the results.

3 The Medical Task

The main goal of medical ImageCLEF task is to improve the retrieval of medical images from heterogeneous and multilingual document collections containing images as well as text. Queries

Language	Experiment	Initial Query	Expansion	Weight	MAP	Rank
Dutch	sinaiNlEnFbOkapiExp1	title + narr	with	Okapi	0.1261	4/4
French	sinaiFrEnFbOkapiExp1	title + narr	with	Okapi	0.1617	5/8
Italian	sinaiItEnFbOkapiExp1	title + narr	with	Okapi	0.1216	13/15
Portuguese	sinaiPtEnFbOkapiExp1	title + narr	with	Okapi	0.0728	7/7
Spanish	sinaiEsEnFbOkapiExp1	title + narr	with	Okapi	0.1849	4/7

Table 3: Summary of results for the others five bilingual adhoc runs

are formulated with sample images and a sort of textual description explaining the research goal. For the medical task, we have used the list of retrieved images by GIFT[1] which was supplied by the organizers of this track.

Last year, our efforts concentrated in manipulating the text descriptions associated with these images and mixing the partial results lists with the GIFT lists [3]. However, this year our experiments focus in preprocessing the collection using Information Gain (IG) in order to improve the quality of results and to automate the tag selection process.

3.1 Preprocessing the Collection

In order to generate the textual collection we have used the ImageCLEFmed.xml file that links collections with their images and annotations. It has external links to the images and the associated annotations in XML files. It contains relative paths, from the root directory, to all the related files.

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. At every case a group of images and an optional annotation is given. Each image is part of a case and has optional associated annotations, which encloses metadata and/or a textual annotation. All of the images and annotations are stored in separate files. ImageCLEFmed.xml only contains the connections between collections, cases, images, and annotations.

The collection annotations are in XML format. The majority of the annotations are in English but a significant number is also in French (in the CASImage collection) and German (in the Pathopic collection), with ew cases not contain any annotation at all. The quality of the texts varies across collections and even within the same collection.

For the MIR subset, specifically designed regular expressions have been applied in order to get different segments of information, due to the lack of predefined XML tags. In this way, information such as identificator string, authors, date and so on has been extracted from within the corpus.

We generate 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 to this image. If there were several images of the same case, then the text was copied several times.

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

3.2 Information Gain and Tag Selection

Last year, almost all tags were used to generate the final corpus. Only those labels that seemed not to provide any information were removed, like the *LANGUAGE* tag. But this year these tags have been selected according to the amount of information theoretically supplied. For this, we have used the information gain measure as a method to select the best tags in the collection.

[1] http://www.gnu.org/software/gift/

The main goal was to determine whether the results obtained from a corpus where tags have been reduced by discarding those with low IG may show higher performance levels. The aim is to eliminate those tags that do not provide further information or that introduce noise, therefore degradating results.

At the beginning, experiments with only 10%, 20%, 30%, ..., 100% of those labels with highest associated IG were performed, using 2005 data for evaluation. Once results were analyzed, most accurated results were obtained with 20%, 30% and 40% of the total of available tags, being these ones the collections used in the submitted experiments for the 2006 campaign.

The method applied consists in computing the information gain for every tag at every subcollection. Since each subcollection (CASImage, Pathopic, Peir and MIR) has a different set of tags, the information gain was calculated using each subcolletion as scope, isolating each one from the others. Let C be the set of cases, E the value set for the E tag, then the formula applied is as follows:

$$IG(C|E) = H(C) - H(C|E) \tag{1}$$

where

$IG(C\|E)$	is the information gain for the E tag,
$H(C)$	is the entropy and
$H(C\|E)$	is the relative entropy

In order to calculate this value, we compute the entropy of the set of cases C as:

$$H(C) = -\sum_{i=1}^{|C|} p(c_i)\log_2 p(c_i) = -\sum_{i=1}^{|C|} \frac{1}{|C|}\log_2 \frac{1}{|C|} = -\log_2 \frac{1}{|C|} \tag{2}$$

And the entropy of the set of cases C conditioned by the tag E would be:

$$H(C|E) = \sum_{j=1}^{|E|} \frac{|C_{e_j}|}{|C|}\left(-\sum_{i=1}^{|C_{e_j}|} \frac{1}{|C_{e_j}|} log_2 \frac{1}{|C_{e_j}|}\right) = -\sum_{j=1}^{|E|} \frac{|C_{e_j}|}{|C|} log_2 \frac{1}{|C_{e_j}|} \tag{3}$$

where

C_{e_j}	is the subset of cases in C having the tag E set to the value e_j (this value is a combination of words where order does not matter)

Therefore, we can conclude the final equation for the computation of the information gain supplied by a given tag E over the set of cases C as follows:

$$IG(C|E) = -\log_2 \frac{1}{|C|} + \sum_{j=1}^{|E|} \frac{|C_{e_j}|}{|C|} log_2 \frac{1}{|C_{e_j}|} \tag{4}$$

For every tag in every collection, its information gain is computed. Then, the tags selected to compose the final collection are those showing high values of IG. Once the document collection was generated, experiments were conducted with the LEMUR[2] retrieval information system, applying the Kl-divergence weighting scheme.

3.3 Experiment Description

Our main goal is to investigate the effectiveness of filtering tags using IG in the text collection. For this, we have accomplished several experiments using the ImageCLEFmed2005 in order to determinate the best tag percentage.

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

Experiment	Precision	Rank
IPAL_Textual_CDW (best result)	0.2646	1
SinaiOnlytL30	0.1178	19
SinaiOnlytL40	0.1133	20
SinaiOnlytL20	0.0990	21

Table 4: Performance of official runs in Medical Image Retrieval (text only)

First, we have carried out experiments with 10%, 20%...100% of tags and we have evaluated the results with the relevance assessments of the 2005 collection. Based on the result obtained, we have only submitted runs with 20%, 30% y 40% of tags for the 2006 collection because these corpus reported the best results. Thus for each experiment, we have submitted 3 runs (one per corpus generated at: 20%, 30% and 40% of all available tags).

We wanted also to compare the obtained results when we only use the text associated to the query topic and the results when we merge visual and textual information. For this, first experiment has been performed as baseline case. This experiment simply consists of taking the text associated to 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.

The remain experiments start from the ranked lists provided by the GIFT tool. The organization provides list of relevant images generated by GIFT for each query. For each list/query we have used an automatic textual query expansion using the associated text to the top ranked images from GIFT lists. Thus, we have added the text associated to the first four images from the GIFT list to the original textual query in order to generate a new textual query. Then, the new textual query is submitted to the LEMUR system and we obtain a new ranked list. Thus, for each original query we have 2 partial lists: one (expanded) text list and one GIFT list. The last step consists of merging these partial resulting lists using some strategy in order to obtain one final list (FL) with relevant images ranked by relevance. The merging process was done given different weight of importance to the visual (VL) and textual lists (TL):

$$FL = VL * \alpha + TL * \beta, \text{ with } \alpha + \beta = 1 \tag{5}$$

In order to set these parameters we have again launched some experiments with the 2005 collection varying α and β in the range [0,1] with step 0.1 (i.e., 0, 0.1, 0.2,...,0.9 and 1). After analyzing the results, we have submitted runs with β set to 0.5, 0.6 and 0.7 for the 2006 collection.

These 3 experiments and the baseline experiment (that only uses textual information of the query) have been accomplished over the 3 different corpus generated with 20%, 30% and 40% of tags. All textual experiments have been carried out with LEMUR using Pseudo Relevance Feedback and the Kl-divergence weighting scheme, as pointed out previously. In summary, we have submitted 12 runs.

3.4 Results

The total runs submitted at ImageCLEFmed2006 for text only were 31 and for mixed retrieval were 37.

Table 4 shows the results for text only retrieval with the SINAI system. Unfortunately, due to a computing processing mistake all our mixed runs obtain the same results than the visual GIFT baseline (0.0467). At the moment of writing of this paper we are modifying our system in order to solve this problem.

4 Conclusions and Further Work

In this paper, we have presented the experiments carried out in our participation in the ImageCLEF campaign.

For the adhoc task, we have tried a new Machine Translation module. The application of some heuristics improves the bilingual results, but it is necessary to study the queries with poorest results, in order to improve them. Our next work will be the improvement of the results in the IR phase, applying new techniques for query expansion (using thesauri or web information) and the investigation in other heuristics for the Machine Translation module.

For the medical task, we have tried to apply Information Gain in order to improve the results. Unfortunately, the performance obtained has been very poor. In addition, for mixed runs our system has a computing mistake and result obtained are no conclusive. However, we consider that the Information Gain is a good idea and a widely used method to filter information without the need of a manual tag selection. Thus, our next step will focus on improving the visual lists and the merging process.

5 Acknowledgements

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

References

[1] Paul Clough, Michael Grubinger, Thomas Deselaers, Allan Hanbury, Henning Müller: Overview of the ImageCLEF 2006 photo retrieval and object annotation tasks. In Proceedings of the Cross Language Evaluation Forum (CLEF 2006), 2006.

[2] Henning Müller, Thomas Deselaers, Thomas Lehmann, Paul Clough, William Hersh: Overview of the ImageCLEFmed 2006 medical retrieval and annotation tasks.In Proceedings of the Cross Language Evaluation Forum (CLEF 2006), 2006.

[3] M.T. Martín-Valdivia, M.T., García-Cumbreras, M.A., Díaz-Galiano, M.C., Ureña-López, L.A., Montejo-Raez, A.: SINAI at ImageCLEF 2005. In Proceedings of the Cross Language Evaluation Forum (CLEF 2005), 2005.