University of Alicante in ImageCLEF2005[*]

Rubén Izquierdo-Beviá, David Tomás, Maximiliano Saiz-Noeda and José Luis Vicedo.
Departamentos de Lenguajes y Sistemas Informáticos. Alicante. Spain
{ruben,dtomas,max,vicedo}@dlsi.ua.es

August 15, 2005

Abstract

This paper describes the participation of the University of Alicante (UA) in CLEF 2005 image retrieval task. For this purpose we used an image retrieval system based on probabilistic information combined with ontological information and a feedback technique. Several information streams are created using different sources: stems, words and bigrams; the final result is obtained combining them. Also a voting-based strategy has been developed joining three different systems of participant Universities.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries

General Terms

Streams, Combination, Ontology, Feedback

Keywords

Text-based image retrieval, Stream combination, Automatic extracted ontology

1 Introduction

Nowadays it's well known the development of the information society and the huge amount of documents it produces. All sorts of documents are generated: plain text, images, videos, source code... This amount of documents makes necessary the use of automatic techniques when trying to access them. Specifically, information retrieval (IR) techniques are related to the task of retrieving relevant documents from user queries and a large set of documents. One subtask of IR is multimedia retrieval, where documents are images, videos, or any kind of multimedia document. Multimedia-based applications are very important in many fields such as medical applications, multimedia databases [1], ...

Within the ImageCLEF task we have developed an image retrieval system. An image retriever is an IR system that recovers relevant images. Mainly, there are two approaches to Image Retrieval [6]. On the one hand we have Content-Based Image Retrieval (CBIR). This approach deals with primitive features of the image using artificial vision techniques. On the other hand there are

[*] This work has been developed in the framework of the project CICYT R2D2 (TIC2003-07158-C04) and it has been partially funded by the Spanish Government through the grant BES-2004-3935.

techniques based on the text that describes the image. Moreover, there are hybrid ones that combine both approaches.

There are two main techniques for text-based systems: probabilistic and natural language processing (NLP) [2]. The former use probabilistic information to make the retrieval while the other exploit the use of NLP techniques such as taggers, parsers, chunkers, named entity recognizers, etc.

Our system combines probabilistic and automatic extracted knowledge from text that describes the image. We have initially used a probabilistic information retrieval system: Xapian [7]. The main idea consists on creating three information streams, each one with different information and combine them to reach the final result.

Furthermore, an ontology has been created using St. Andrews Corpus [8]. The idea is to know the category related to a query. In this way we can select from the final retrieved list those documents with the same category than the query.

In order to deal with the multilingual view we have made an analysis with the ImageCLEF 2004 queries set and several automatic translators. Topics in different languages are first translated into English and then retrieval is performed.

This paper is structured as follows. Section 2 presents the system in depth. Section 3 explains the development of the image ontology. Section 4 provides the result of our participation in ImageCLEF 2005. Section 5 describes the joint participation. Some conclusions and future work close the document.

2 System

The system implements an approach based on probabilistic information and knowledge extracted from the corpus. The system uses Xapian [7], a probabilistic and boolean information retrieval system and the process is divided into two phases: indexing and retrieval. During the indexing phase three indexes with different information are created. When we have a query, we make a retrieval in each index and then combine the results to get the final goal. Figure 1 shows this process.

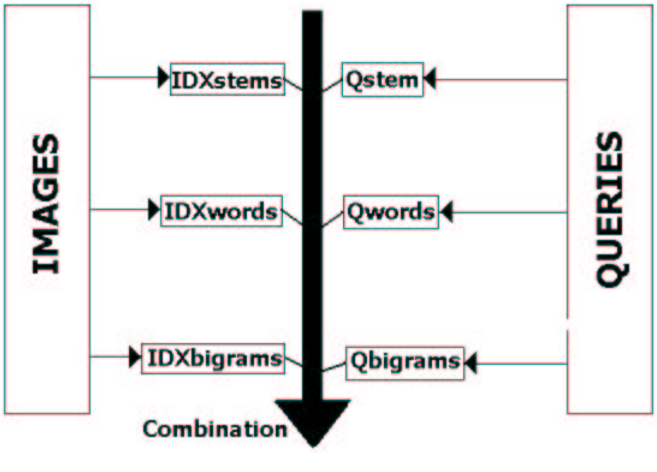
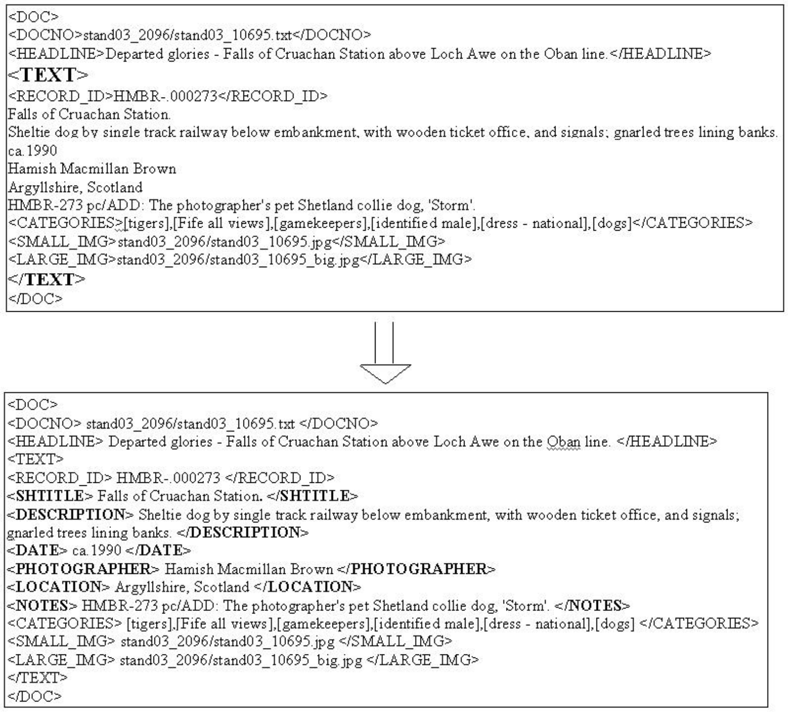


Figure 1: Generation of streams

2.1 Indexing

In this step we create three indexes, each one with different information: words, stems and stem bigrams. The first to do is to process the image text to extract relevant information. All the text that follows the image is relevant, but each field will have a weight depending on its relevance. This is done by means of a set of manually developed patterns, which are regular expressions proposed in the image retrieval system presented at the ImageCLEF 2004 campaign [5]. In this way we have more information and can distinguish between useful and useless information. Figure 2 shows an example of the use of the patterns. From the <TEXT> field and by means of the patterns the fields extracted are: *shtitle, description, date, photographer, location and notes.*



```
<DOC>
<DOCNO>stand03_2096/stand03_10695.txt</DOCNO>
<HEADLINE>Departed glories - Falls of Cruachan Station above Loch Awe on the Oban line.</HEADLINE>
<TEXT>
<RECORD_ID>HMBR-.000273</RECORD_ID>
Falls of Cruachan Station.
Sheltie dog by single track railway below embankment, with wooden ticket office, and signals; gnarled trees lining banks.
ca.1990
Hamish Macmillan Brown
Argyllshire, Scotland
HMBR-273 pc/ADD: The photographer's pet Shetland collie dog, 'Storm'.
<CATEGORIES>[tigers],[Fife all views],[gamekeepers],[identified male],[dress - national],[dogs]</CATEGORIES>
<SMALL_IMG>stand03_2096/stand03_10695.jpg</SMALL_IMG>
<LARGE_IMG>stand03_2096/stand03_10695_big.jpg</LARGE_IMG>
</TEXT>
</DOC>
```

```
<DOC>
<DOCNO> stand03_2096/stand03_10695.txt </DOCNO>
<HEADLINE> Departed glories - Falls of Cruachan Station above Loch Awe on the Oban line. </HEADLINE>
<TEXT>
<RECORD_ID> HMBR-.000273 </RECORD_ID>
<SHTITLE> Falls of Cruachan Station. </SHTITLE>
<DESCRIPTION> Sheltie dog by single track railway below embankment, with wooden ticket office, and signals;
gnarled trees lining banks. </DESCRIPTION>
<DATE> ca.1990 </DATE>
<PHOTOGRAPHER> Hamish Macmillan Brown </PHOTOGRAPHER>
<LOCATION> Argyllshire, Scotland </LOCATION>
<NOTES> HMBR-273 pc/ADD: The photographer's pet Shetland collie dog, 'Storm'. </NOTES>
<CATEGORIES> [tigers],[Fife all views],[gamekeepers],[identified male],[dress - national],[dogs] </CATEGORIES>
<SMALL_IMG> stand03_2096/stand03_10695.jpg </SMALL_IMG>
<LARGE_IMG> stand03_2096/stand03_10695_big.jpg </LARGE_IMG>
</TEXT>
</DOC>
```

Figure 2: Example of pattern application

With the information of the fields, and depending on the upper or lowercase in the first letter of the token (word, stem or bigram), a weight to each of this tokens is assigned[1](see Table 1).

FIELD	WEIGHT
Headline	5
Shtitle	4
Description	1
Data	3
Photo	3
Location	3
Notes	0
Categories	8

Table 1: Weights assigned to each field in image file

Finally, the weight assigned to each token is:

$$W_{token} = \begin{cases} 100 * field\_weight & \text{if 1st letter is uppercase} \\ 50 * field\_weight & \text{if 1st letter is lowercase} \end{cases}$$

[1] The weights have been adjusted to have the best performance over Image CLEF 2004 query set.

Using this weighting scheme, we create three indexes and the indexing phase finishes.

2.2 Retrieval

In this phase the query topics are processed and a list of relevant documents to each query is given. The main idea is to perform three independent retrievals, one in each index and then combine them to get a unique list of retrieved images.

The first step prepares the query to be posed to the retrieval system. Given a topic, stop words are removed and then each word is transformed to have three queries, one for each index.

Once we have the three queries, a first retrieval in each index is performed. The second step is applies the relevance feedback. Xapian allows us to use feedback in a simple way: we must only select those documents which we consider relevant. Some experiments over ImageCLEF 2004 query set reveal that the number of documents to get the better results is twenty three.

In the next step the ontology is used. With this ontology we know both the categories related to the query and the categories of each retrieved document. Using this information the relevance of the documents having any category in common with the query can be increased. This process is done separately in the three retrieved lists.

The last step combines the three lists to get a global result. We believe that each stream gives a different kind of information, and thus, each stream must have a different weight. We made an analysis to make the best weight tuning considering the contribution of each information flow.

- Stem flow: It can increase the recall because with the stemming, same words but morphologically different can be matched.
- Word flow: It can increase the precision. Only words exactly equal are matched.
- Bigram flow: It can increase the precision. Word pairs, such as compound noun phrases.

So, the weights selected to do the combination are[2]:

- Stem flow: 0.5
- Word flow: 0.1
- Bigram flow: 0.3

In the combination, each document has as score the sum of its flow scores multiplied by their corresponding weight.

$$W_{Doc} = 0.5 * W_{Flow} + 0.1 * W_{Word} + 0.3 * W_{Bigram}$$

2.3 Multilingual view

In order to deal with multilingual features we have decided to make an analysis of several online translators. Among different well-known translators such as Babel [9], Reverso [12], WordLingo [13], Epals [10] and Prompt [11], WordLingo was selected due to its the best average result for the treated languages according to the ImageCLEF 2004 query set and the St. Andrews corpus. The test languages for the system are Dutch, French, German, Greek, Italian, Japanese, Portuguese, Russian, Chinese and Spanish.

3 The image ontology

In this section we are going to describe how the ontology is created and how the system uses this information in the retrieval process. This ontology represents a knowledge base, and it is an approach to use knowledge resources in information retrieval. The idea is to extract the categories related with a query and then retrieve documents with the same categories.

[2] These weights have been adjusted doing an analysis over ImageCLEF 2004 topics to get the best performance.

3.1 Ontology building

The ontology has been automatically built using the St. Andrews corpus. Each image caption has a field called <CATEGORIES> (see figure 2). For each image in the corpus the categories and the words of the <TITLE> field are extracted. The ontology is used to create an “ontology-document” database that relate each category with the corresponding words. The process is shown in figure 3.

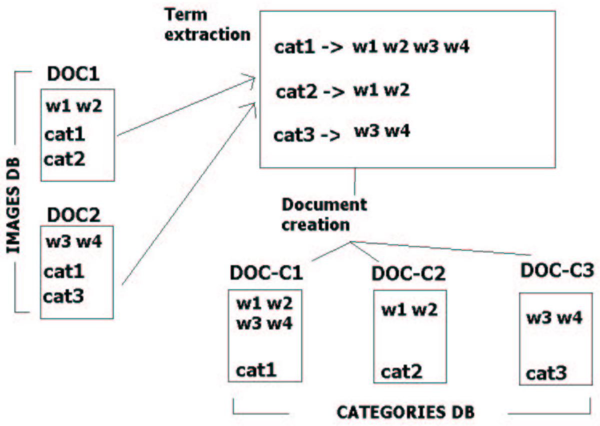
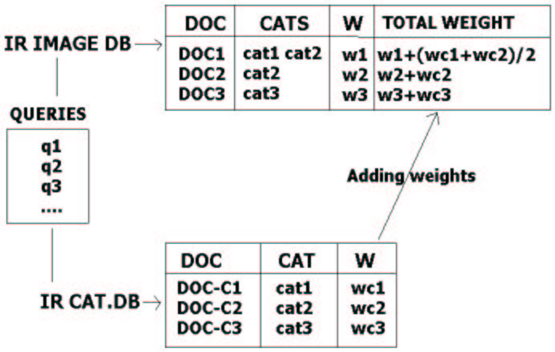


Figure 3: Creation of the ontology

3.2 Ontology use

Queries are posed to the ontology index in order to retrieve a ranked list of “ontology-documents”. The score of the category document ’C’ indicates the probability of a query to require a document of this category ’C’. The category ranking results will be used for document re-ranking using the lists of relevant documents (one for each stream) obtained from first step image retrieval process. When we make the retrieval in each of the index created (stem, word and bigram), we have a list of documents. Each of these documents has several categories, so if we know the categories related to the query, we can increase the weight of those documents having some category in common with the query. Figure 4 shows this process, where weights extracted from the category database (wc1, wc2 and wc3) are added to the total weight of the image according to their common categories.



DOC	CATS	W	TOTAL WEIGHT
DOC1	cat1 cat2	w1	w1+(wc1+wc2)/2
DOC2	cat2	w2	w2+wc2
DOC3	cat3	w3	w3+wc3

DOC	CAT	W
DOC-C1	cat1	wc1
DOC-C2	cat2	wc2
DOC-C3	cat3	wc3

Figure 4: Using the ontology

4 Experiments and Results

Four experiments have been proposed for the evaluation. In each one we combine different techniques in order to discover how these techniques help in the retrieval process. The features that have been merged in different experiments are:

- Uses of stems, words or bigrams(tokens) adjusting their weights.
- Use of correct fields adjusting their weights.
- Selection of different flow weights when combining the results.
- Use of categories (ontology).
- Use of automatic feedback.

Combining all these features, over 100 experiments have been performed, but we have only selected the best four in order to participate in ImageCLEF 2005. Experiments selected and their features are shown in next table:

	STEM	WORD	BIGRAM	CATS.	FEEDBACK
Baseline	X				
Experiment1	X	X	X		
Experiment2	X	X	X		X
Experiment3	X	X	X	X	X

Table 2: Feature selection for each retrieval experiment.

The results of the monolingual system over ImageCLEF05 query set are:

EXPERIMENT	MAP
Baseline	0.3944
Experiment1	0.3942
Experiment2	0.3909
Experiment3	0.3966

Table 3: Results for UA monolingual adhoc retrieval in ImageCLEF 2005.

As we can see in the previous table, Experiment3 provides the best performance. Experiment3 uses stems, words and bigrams, and also implements feedback and categories knowledge.

According to Baseline, Experiment1 and Experiment2 comparison it seems that word and bigram adding do not improive the results[3].

The improvement related to he use of the knowledge ontology can observed when comparing Experiment2 and Experiment3. This increase could be greater with a more sophisticated ontology, for example, use a more general ontology that combines more general concepts and specific domain terms.

[3] During the experiments over ImageCLEF 2004 dataset, word and bigram adding provided a slight improvement.

5 Joint participation

Apart from the system described, we have made a joint participation within the R2D2 project[4] framework integrating our system and the ones belonging to UNED group from Madrid and SINAI group from Jaén. We have developed a voting among them. A combination between UA and SINAI has been done for English, Dutch, French, German, Italian, Russian and Spanish. The three systems have been only combined for Spanish. The systems selected for the combination implement feedback and use query titles and automatic translation.

The voting has been developed using the weights of each document in each retrieved list. The first step makes a normalization of the weights for each document list dividing each weight by the maximum weight of all the documents in the list. Then the different weights of each document in each list are summed. In this way, documents that appear in the three lists are more relevant than those appearing in just two lists. Finally the documents are sorted according to the calculated weight and a final document list is generated.

MAP results obtained from the collaborative experiments are shown in the Table 4. Ranking positions are shown in brackets.

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

Table 4: Comparing results in voting-based collaborative system.

It can be observed that this voting system improves the results of the individual separated systems in the most of the languages: English, Dutch, German, Italian and Spanish (lat).

6 Conclusion and future work

This paper has presented the approach used by the University of Alicante in the ImageCLEF 2005 adhoc retrieval task. This image retrieval system implements a combination of probabilistic and knowledge-based approaches. The implementation and tuning of the method have been shown and results have been explained.

To continue improving the system there are several ways that can be taken in account. One of them is to consider the use of NLP to improve the information retrieval [4]. For example, a chunker or parser or even better a named entity recognizer can be used to detect noun phrases or entities. In this way we could create a noun phrase stream to be combined with the existing ones.

Another work to be developed is the creation and management of the ontology, that is, the use of knowledge in the retrieval process [3]. A more complex ontology can be created considering the possibility of changing the knowledge representation.

[4] R2D2 project: Question Answering in Digital Documents. Reference TIC2003-07158-C04 financed by the Science Technology Department 2003-2006 of the Spanish Government. The main goal of the project is the evaluation and development of Question Answering and Document Retrieval systems in multilingual scenarios. See http://gplsi.dlsi.ua.es/r2d2/indexEn for more details.

References

[1] Databases Forsyth Computer. Benchmarks for storage and retrieval in multimedia.

[2] A. Goodrum. Image information retrieval: An overview of current research, 2000.

[3] V. Kashyap. Design and creation of ontologies for environmental information retrieval, proceedings of the 12th workshop on knowledge acquisition, modeling and management (kaw'99), banff, canada, october 1999. In *KAW'99 Conference*, 1999.

[4] David D. Lewis and Karen Sparck Jones. Natural language processing for information retrieval. *Communications of the ACM*, 39(1):92–101, 1996.

[5] Maximiliano Saiz-Noeda, José Luis Vicedo and Rubén Izquierdo. Pattern-based Image Retrieval with Constraints and Preferences on ImageCLEF 2004. In *Working Notes for the CLEF 2004 WorkShop*, Bath, United Kingdom, 2004.

[6] Paul Clough and Mark Sanderson and Henning Müller. The CLEF Cross Language Image Retrieval Track (imageCLEF) 2004. In *Working Notes for the CLEF 2004 WorkShop*, Bath, United Kingdom, 2004.

[7] www. The xapian project. http://www.xapian.org/.

[8] www. St Andrews University Library photographic collection. http://specialcollections.st-and.ac.uk/photcol.htm, 2004.

[9] www. BabelFish translator. http://world.altavista.com/, 2005.

[10] www. Epals translator. http://www.epals.com/translation/translation.e, 2005.

[11] www. Prompt translator. http://translation2.paralink.com/, 2005.

[12] www. Reverso translator. http://www.reverso.net/, 2005.

[13] www. WordLingo translator. http://www.worldlingo.com/en/products_services/worldlingo_translator.html, 2005.