A Probabilistic, Text and Knowledge-Based Image Retrieval System

Rubén Izquierdo-Beviá, David Tomás, Maximiliano Saiz-Noeda, and José Luis Vicedo

Departamento de Lenguajes y Sistemas Informáticos.
Universidad de Alicante. Spain
{ruben, dtomas, max, vicedo}@dlsi.ua.es

Abstract. This paper describes the development of an image retrieval system that combines probabilistic and ontological information[1]. The process is divided in two different stages: indexing and retrieval. Three information flows have been created with different kind of information each one: word forms, stems and stemmed bigrams. The final result combines the results obtained in the three streams. Knowledge is added to the system by means of an ontology created automatically from the St. Andrews Corpus. The system has been evaluated at CLEF05 image retrieval task.

1 Introduction

An image retriever is an IR system that discovers relevant images. Mainly, there are two approaches to Image Retrieval [1]. On the one hand we have Content-Based Image Retrieval (CBIR). This approach deals with primitive features of the image using computer vision techniques. On the other hand there are techniques based on the text that describes the image. Moreover, there are hybrid ones that combine both approaches.

Our system combines probabilistic and automatic extracted knowledge from the text that describes the image. We have initially used a probabilistic information retrieval system: Xapian[2]. The knowledge is incorporated using an ontology created automatically from the St. Andrews Corpus.

2 The System

Our system relies on Xapian, a probabilistic and boolean information retrieval system. The process is divided in two stages: indexing and retrieval.

2.1 Indexing

In this stage, we process the text of the image and create three indexes using words, stems and stemmed bigrams. The text is analyzed by means of a set of

[1] This work has been partially supported by the Spanish Government (CICYT) with grant TIC2003-07158-c04-01.

[2] The Xapian Project, http://www.xapian.org

patterns and several fields are extracted from it. We assign a weight to each field, depending on the relevance of the information contained on it. The fields extracted and the weights selected are shown in table 1.

Table 1. Weights assigned to each field in the image file

FIELD	Headline	Short title	Description	Data	Photographer	Location	Notes
WEIGHT	5	4	1	3	3	0	8

For each image we create a document to be indexed. This document consists of weighted tokens extracted from the text that describes the image. Tokens can be words, stems and stemmed bigrams. In this way we create three indexes using different tokens. The weight assigned to each token is:

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

2.2 Retrieval

In the retrieval stage, for each query topic we make three retrievals (one for each index) and combine the results to get a single list of ranked documents.

The first step prepares the query to be processed by the retrieval system. Stop words are removed and words are processed to obtain stems and stemmed bigrams. The retrieval process can be summarized in these steps:

1. Retrieval in the corresponding index
2. Apply relevance feedback to expand the query[3]
3. Retrieve with the expanded query
4. Enrich the results with the ontology information

As a result we obtain three document lists, one for each index. The next step is to combine them to get the final result: a single list of weighted documents. Each information stream provides a different kind of information, and thus, each stream must have a different weight. We analyzed the system's performance to obtain the best weight tuning considering the contribution of each information flow. The weights assigned to stem, word and bigram flows are: 0.5, 0.1 and 0.3, respectively. When combining, each document is scored by the sum of its flow

[3] Xapian allows us to apply relevance feedback by selecting a number of documents considered relevant. We have selected the first twenty three documents due to some experiments over the ImageCLEF 2004 query set reveal that this is the number of documents suitable to get the best results.

scores multiplied by their corresponding weight ($0.5 * W_{Flow} + 0.1 * W_{Word} + 0.3 * W_{Bigram}$).

3 Multilingual View

We have used an automatic online translator to deal with multilingual features. The process consists on translating the query topics into English and then use the monolingual system described in the previous section. We compared several translators in order to select the best performing one. This analysis was carried out using the ImageCLEF2004 query set and the St. Andrews Corpus. The translators reviewed were Babel[4], Reverso[5], WordLingo[6], Epals[7] and Prompt[8]. The best performance was achieved by WordLingo.

4 Ontology

The ontology has been created automatically from the St. Andrews Corpus. Each image in this corpus has a field called <CATEGORIES>. We can extract the words contained in the rest of the fields and match them with these categories. In this way, we created an ontology, where each category is related to the images belonging to it through the words that describe these images (category descriptor vector).

The ontology is used as follows: the system computes the similarity between the query and the categories using the category descriptor vectors, and the weight obtained boosts document similarity in the relevant document lists previously obtained in the retrieval stage. This way, the relevance of documents having any category in common with relevant categories is increased according to the relevance of the category obtained.

5 Experiments and Results

Four experiments have been carried out combining different features and techniques. The features merged in the different experiments are: the kind of tokens used (stem, words, bigrams), the fields selected and their weights, the weights for flow combination, the use of ontology and the use of automatic feedback.

With these features we developed over 100 experiments. The characteristics and results of the best ones are shown in table 2.

As shown, *Experiment3* provides the best performance. It uses stems, words and bigrams implementing feedback and category knowledge. Stream combination and ontology information improve the overall performance.

[4] http://world.altavista.com/

[5] http://www.reverso.net/

[6] http://www.worldlingo.com/en/products_services/worldlingo_translator.html

[7] http://www.epals.com/translation/translation.e

[8] http://translation2.paralink.com/

Table 2. Feature selection for each retrieval experiment

	STEM	WORD	BIGRAM	CATS.	FEEDBACK	MAP
Baseline	X					0.3944
Experiment1	X	X	X			0.3942
Experiment2	X	X	X		X	0.3909
Experiment3	X	X	X	X	X	0.3966

6 Conclusions and Future Work

In this paper we have presented an image retrieval method based on probabilistic and knowledge information. The system implements a text-based multimedia retrieval system. We have used Xapian, a probabilistic and boolean information retrieval system, and an ontology created automatically from the St. Andrews Corpus.

We can conclude that our system has reached a high performance with a simple idea: the combination of different information streams and the use of knowledge.

Having in mind CLEF05 competition [2] and comparing our results with other participant systems, our system performs better than CBIR (visual retrieval) approaches and our results are also above the average MAP for different features combination in text-based systems. Our best result (*Experiment3*) reached 0.3966 for English, taking into account that the average MAP for English runs is 0.2084. *Experiment3* implements feedback, while the average MAP for runs using feedback is 0.2399. Finally, we used only title as query, with the average MAP for runs using title being 0.2140.

The system can be improved in different ways. First consider the use of NLP to improve the information retrieval [3]. Another task to be developed is the creation and management of the ontology, that is, the use of knowledge in the retrieval process [4].

References

1. Paul Clough and Mark Sanderson and Henning Muller: The CLEF Cross Language Image Retrieval Track (imageCLEF) 2004. In: Working Notes for the CLEF 2004 WorkShop, Bath, United Kingdom (2004)
2. Paul Clough, Henning Muller, Thomas Deselaers, Michael Grubinger, Thomas M. Lehmann, Jeery Jensen, William Hersh: The CLEF 2005 Cross-Language Image Retrieval Track. In: Proceedings of the Cross Language Evaluation Forum 2005, Springer Lecture Notes in Computer science, Viena, Austria (2006 (to appear))
3. Lewis, D.D., Jones, K.S.: Natural language processing for information retrieval. Communications of the ACM 39(1) (1996) 92101
4. Kashyap, V.: Design and creation of ontologies for environmental information retrieval, proceedings of the 12th workshop on knowledge acquisition, modeling and management (kaw'99), ban, canada, october 1999. In: KAW'99 Conference. (1999)