Information retrieval of visual descriptions with IR-n system based on passages

Sergio Navarro, Fernando Llopis, Rafael Muñoz Guillena, Elisa Noguera
Grupo de Investigación en Procesamiento del Lenguaje Natural y Sistemas de Información
Departamento de Lenguajes y Sistemas Informáticos
University of Alicante, Spain
snavarro,llopis,rafael,elisa@dlsi.ua.es,

Abstract

This paper describes an approach made to the development of a textual image retrieval system, by the university of Alicante using IR-n, a Information Retrieval (IR) text-based system. With only a minimal quantity of adaptations to the features of this task, our system has obtained precision results over the mean average of participants at ImageCLEF07: for English (0.1604 vs 0.1388) and for Spanish (0.1482 vs 0.1450). For German, our results were under the mean (0.0991 vs 0.1331), it could be due to our system does not incorporate a splitter for the treatment of this agglutinative language. We obtain these results, without incorporate specific adaptations of the dominion of the recovery of images. This leads us to believe that we start from a good base point from which to work to obtain better results.

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; H.2.3 [**Database Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Question answering, Questions beyond factoids

1 Introduction

Nowadays the amount of information documents that society produces is huge. All sort of documents are generated: text-planes, images, videos, source codes, etc. This amount of documents requires the use of automatic techniques for its access.

In order to take up this challenge, systems of information retrieval are used (IR). The primary target of these systems, is to locate the relevant documents from a user query in a document collection.

In order to attend and to stimulate the development of systems of IR based on the different European languages, the campaigns CLEF have been created. These are annual competitions where different systems from IR compete. Specifically, within the scope of the CLEF [2], an area specialized in the images retrieval exists, wich is the ImageCLEF [1] [5].

For our task of the ImageCLEF we have used the IR, IR-n [7]. It's a information retraival system that using statistical techniques, has given good results in flat text based tasks. The objective to use a system of these characteristics is to contrast, in the scope of the recovery of images, the results of a statistical system, with others which includes NLP techniques.

This paper is structured as follows: Firstly, presents the main characteristics of the IR-n system; the following section explains the task with which we have evaluated the system and the carried out training; finally, in the last section we present the results and the conclusions.

2 IR-n System

For this approach we have used IR-n, that is a IR based on passages. The RP systems, consider each document like a set of passages, where a passage defines as a portion or contiguous text block. This kind of systems opposite to those that are based on documents, allow to consider the proximity of appearance of the words in a document, to value their relevance [7].

IR-n like RP system difference of other systems pertaining to the same category, in the method proposed for the definition of the passages. The unit that uses the IR-n system to define the passages is the phrase. Thus, the passages are defined by a number of consecutive phrases of the document. [7].

In this section the main characteristics of the IR-n system are described, and the techniques used for ImageCLEF 2007 are detailed.

2.1 Resources: stemmers and stopword list

Stemmers and stopword list are used with the objective to discriminate the information that is going to be used for retrieval. Thus, the stopword list of each language contains those words that, in spite of appearing in the query, does not consider important its appearance in the document to determine if this one is relevant or no. As far as stemers, these, obtain the root of a word eliminating the suffixes and prefixes of the same ones, for their indexing and search.

IR-n uses stemmers and the stopword list available in the web `www.unine.ch/info/clef`.

2.2 Weighting models

The weighting models allow to quantify the similarity between a text (a complete document or a passage) and the query. These measures are based fundamentally on the terms that share the text and the query, as well as on the discriminatory importance of each term. IR-n uses several weighting models. For this competition we have used dfr [3] and okapi [6]. The document ranking produced by each weighting model is obtained using the same general expression, this one is defined as the product of the weight of a term in the document by the weight of the term in the query.

$$sim(q,d) = \sum_{t \in q \wedge p} w_{t,p} \cdot w_{t,q} \tag{1}$$

Variables List Here is described the list of variables used in the following formulas. 2, 3:

- $f_{t,p}$ is the frequency of the term t on the passage p,
- $f_{t,q}$ is the frequency of the term t on the query q,
- n is the number of documents in the collection,
- n_t is the number of documents in which it appears t,
- c, k_1, b and k_3 are constant values,

- ld is the length of the document,
- avg_{ld} is the average of the length of the documents

Okapi Using the okapi model, the weight of a passage p for a query q is given by:

$$\begin{aligned} w_{t,p} &= \frac{(k_1+1)\cdot f_{t,p}}{K\cdot f_{t,p}} \\ w_{t,q} &= \frac{(k_3+1)\cdot f_{t,q}}{k_3\cdot f_{t,q}}\cdot w_t \\ K &= (1-b)+b\cdot\frac{ld}{avr_ld} \\ w_t &= log_2\frac{n-n_t+0.5}{n_t+0.5} \end{aligned} \tag{2}$$

DFR Using this model, the weight of a passage p for query q is given by:

$$\begin{aligned} w_{t,q} &= f_{t,q} \\ w_{t,p} = (log_2(1+w_t)+w'_{t,p}\cdot log_2(\frac{1+w_t}{w_t}))\cdot &\frac{f_t+1}{n_t\cdot(w'_{t,p}+1)} \\ w'_{t,p} &= f_{t,p}\cdot log_2(1+\frac{c\cdot avr_{ld}}{ld}) \\ w_t &= \frac{f_t}{n} \end{aligned} \tag{3}$$

2.3 Query expansion

Most IR systems use query expansion techniques [4] based on adding the most frequent terms contained in the most relevant documents to the original query. The IR-n architecture allows us to use query expansion based on either the most relevant passages or the most relevant documents. In previous researches, we obtained better results using the most relevant passages.

3 Training

IR-n is a parametrizable system, that allows to adapt it to the concrete characteristics of the task to make. The parameters for his configuration are: the number of phrases that form a passage, the weighting model to use, the type of expansion, the number of documents/passages on which the expansion is based, and the average number of words by document.

This section describes the training process which has been carried out in order to obtain the best features to improve the performance of the system. Firstly, the collections and resources are described, and in following section, the specific experiments carried out.

3.1 Data Collections

We have participated in the following monolingual tasks of ImageCLEF 2007: English, German and Spanish. For its training with English and German we have been based on corpus of previous years. Table 3.1 shows the characteristics of the language collections.

- **WDAvg**: is the average of words by document.
- **NrQue**: is the number of queries that are used in the experiments on each collection.

Language	Colection	NDocs	WDAvg	NrQue
English	St Andrews(ImageCLEF2004)	28.133	48	25
English	IAPR TC-12 (ImageCLEF2006)	20.000	40,29	60
German	IAPR TC-12 (ImageCLEF2006)	20.000	34,61	60

Table 1: Data Collections

To the Spanish we did not have previous corpus, that's the reason why we have used as guide the results obtained for the English and German.

As we can see at Table, the difference between corpus of the year 2004 and corpus of the year 2006 is substantial, since the one of the 2004 presents greater amount of information (greater number of documents and greater average of words by document). Furthermore, at corpus of 2006 only a 70% of the documents has all the information, the 10% hasn't description, another 10% has only location and date, and finally last 10% hasn't any annotation. All of this, increases the possibilities of success in a task of textual information retrieval over the corpus of 2004.

The collections of data has a semi structured format. We took advantage of it selecting the information used as input of the IR-n system.

The information that is not interesting for a textual recovery is rejected. Thus, we used as input for IR-n only the fields that correspond with the title (TITLE), the description (DESCRIPTION), notes (NOTES), the place (LOCATION) and the date of the photo (DATES).

The queries associated to each corpus (60 queries by corpus) also has a semi structured format as is possible to see in Figure 1, and in addition the text contains images with similar contents to the target to retrieve.

```
<num> Number: 1 </num>
<title> accommodation with swimming pool </title>
<narr> Relevant images will show the building of an accommodation facility
(e.g. hotels, hostels, etc.) with a swimming pool.
Pictures without swimming pools or without buildings are not relevant.
</narr>
<image> images/03/3793.jpg </image>
<image> images/06/6321.jpg </image>
<image> images/06/6395.jpg </image>
</top>
```

Figure 1: Queries from set of queries of ImageCLEF 2006

Only the queries in English of ImageCLEF 2006 and ImageCLEF 2004 contain narrative (NARR) that accompanies the query, the rest of sets of queries of ImageCLEF 2006 in other languages only contains title (TITLE) as textual information.

Furthermore from a set of 60 queries in ImageCLEF 2006, only 30 were responded with textual information, 10 with visual information and 20 indifferently with textual or visual information.

3.2 Experiments

The aim of the experiment phase is set up the optimum value of the input parameters for each collection. Next, we describe the input parameter of the system:

- **Size of the Passage (sp)**: Number of phrases that form the passage.
- **Weight model (wm)**: We used two weighting models : **okapi** y **dfr**.
- **Opaki parameters**: k_1, b and avg_{ld} (k_3 is fixed as 1000).

- **Dfr parameters**: c and avg_{ld}.
- **Query expansion parameters**: If **exp** has value 1, this denotes we use relevance feedback based on passages in this experiment. But, if **exp** has value 2, the relevance feedback is based on documents. Moreover, **num** denote the number of passages or documents that the expansion will use, and **term** indicates the k terms extracted from the best ranked passages or documents from the original query
- **Evaluation Measure**: MAP Mean average precision (**avgP**) is the evaluation measure used in order to evaluate the experiments.

3.2.1 English

As we can see for corpus of 2004 at Table 2, expansion based on passages and dfr obtains better results. It is important to stand out that, better results are obtained establishing the average length (in bytes) by document to values greater than corpus has.

sp	wm	c	avgld	k1	b	exp	num	term	avgP
5	dfr	4	1600						0.4752
5	**dfr**	**9.5**	**1700**			**1**	**10**	**5**	**0.5128**
5	okapi		50	1	0.2				0.4752
5	okapi		350	3	0.2	1	5	5	0.5086

Table 2: English 2004 Best Results

For 2006 English corpus we can see that the precision is reduced considerably. This is justified by the fact that corpus has minor amount of textual information that in the edition of the 2004, as reflects Table 3.1. And too because corpus of the 2005 has a 30% of incomplete documents. Also we observed, that the best results are obtained with dfr and techniques of expansion based on documents. For this corpus we see that the average length by document, has been standardized, corresponding with numbers nearer the real ones.

sp	wm	c	avgld	k1	b	exp	num	term	avgP
2	dfr	5.5	85						0.1926
3	**dfr**	**8**	**85**			**2**	**5**	**5**	**0.2059**
2	okapi		90	4	0.8				0.1799
2	okapi		1900	4	0.8				0.1800
5	okapi		90	2	0.8	1	10	10	0.1992

Table 3: English 2006 Best Results

3.2.2 German

The results in German are lower as we can see at Table 4.

It could be a combination of two causes: on the one hand the fact that IR-n does not incorporate a mechanism for the treatment of compound words (in spite of being very common in this language), on the other hand the circumstance that the queries in German do not contain narrative, which reduce the success possibilities.

For German the weighting model that better results gives back is okapi, with expansion by documents.

sp	wm	c	avgld	k1	b	exp	num	term	avgP
3	dfr	2	90						0.1487
3	dfr	2	85			1	5	5	0.1742
3	okapi		85	2	0.8				0.1492
3	**okapi**		**85**	**2**	**0.8**	**2**	**5**	**5**	**0.1857**

Table 4: German 2006 Best Results

3.2.3 Experiments Summary

The obtained training results and the results obtained by the participants of the respective ImageCLEFs, are compared at Table 5

Competition	avgP IR-n	avgP ImageCLEF	avgP Best
ImageCLEF04 English	0.5128	0.4155	0.58
ImageCLEF06 English	0.2059	0.152	0.385
ImageCLEF06 German	0.1857	0.121	0.311

Table 5: Compared Results

These results were obtained with the best configuration for each corpus and language. For our participation at ImageCLEF07 we used the same configuration than the best one used with ImageCLEF06 corpus. The decission is justified by the fact that the corpus for ImageCLEF07 is the same that we used at ImageCLEF06, and it will improve our success rate. Despite the fact that we have discarded the 2004 results for the training process, they have helped us to meassure how the features of the corpus can affect the results.

4 Results at ImageCLEFPhoto-2007

Table 6 shows the configurations used for ImageCLEF 2007, and the comparative of the obtained results with the average of all participants by language at monolingual task. For English and German, we used the configurations which obtain the best rsults at training phase using the ImageCLEF06 corpus. For Spanish, we used the same configuration than for English.

lang	sp	wm	c	avgld	k1	b	exp	num	term	avgP	ImgCLEF07
Eng	3	dfr	8	85			2	5	5	0.1604	
	3	dfr	8	85			0	0	0	0.1453	0.1388
Spa	3	dfr	8	85			2	5	5	0.1482	
	3	dfr	8	85			0	0	0	0.1367	0.1450
Ger	3	okapi		85	2	0.8	2	5	5	0.0991	
	3	okapi		85	2	0.8	0	0	0	0.0911	0.1331

Table 6: ImageCLEFPhoto 2007 official average results. Monolingual tasks.

5 Conclusion and future work

In our first ImageCLEF participation, we used an IR text-based system. We used it with a minimal quantity of adaptations to the features of this task. We highlight that its precision results are above average for English and Spanish. The lower results in German could be due to IR-n not

incorporate a mechanism for the treatment of compound words (in spite of being very common in this language). We will work to solve this in future projects.

In addition, analysing the results of training with the corpus of previous years, makes us meassure how the no completeness of the corpus and the absence of theme tags (like in 2004 corpus) result in decrease of precision values, and moreover the reduction of the length of the corpus has a direct effect over the passage size.

Finally, the fact to obtain these results, without to have incorporated specific adaptations of the dominion of the recovery of images, makes us to think, that we start from a good base point, from which work, to obtain better results.

To continue improving the system there are several ways that can be taken in account. One of them is to consider to do an approach to the resolution of a multilingual corpus using the calculation of the documentary relevance in two steps [8]. Another work to be developed is to improve the system with the incorporation of a system CBIR, that complements the textual information retrieval that IR-n carries out. Therefore, as future work we will explore the possibility of shape extraction from images, establishing a relationship between them and associated terms. And finally, we will try to add NLP techniques to the local query expansion.

6 Acknowledgements

This research has been partially funded by the Spanish Government under project TEXT-MESS (TIN-2006-15265-C06-01), and by European Union (EU) under QALL-ME project (FP6-IST-033860), and by the Valencia Government under project number GV06-161.

References

[1] http://ir.shef.ac.uk/imageclef.

[2] http://www.clef-campaign.org.

[3] G. Amati and C. J. Van Rijsbergen. Probabilistic Models of information retrieval based on measuring the divergence from randomness. *ACM TOIS*, 20(4):357–389, 2002.

[4] Aitao Chen and Fredric C. Gey. Combining Query Translation and Document Translation in Cross-Language Retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and et al., editors, *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Lecture notes in Computer Science*, Lecture notes in Computer Science, Trondheim, Norway, 2003. Springer-Verlag.

[5] Michael Grubinger, Paul Clough, Allan Hanbury, and Henning Müller. Overview of the ImageCLEFphoto 2007 photographic retrieval task. In *Working Notes of the 2007 CLEF Workshop*, Budapest, Hungary, September 2007.

[6] Savoy J. Fusion of Probabilistic Models for Effective Monolingual Retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler, and et al., editors, *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003*, Lecture notes in Computer Science, Trondheim, Norway, 2003. Springer-Verlag.

[7] Fernando Llopis. IR-n: Un Sistema de Recuperación de Información Basado en Pasajes. In *PhD thesis*, 2003.

[8] Fernando Martínez Santiago. El problema de la fusión de colecciones en la recuperación de información multilingüe y distribuida: cálculo de la relevancia documental en dos pasos. In *PhD thesis*, 2004.