UJM at ImageCLEFwiki 2008

Christophe Moulin, Cécile Barat, Mathias Géry, Christophe.Ducottet, Christine Largeron
Université de Lyon - UMR 5516
Saint-Étienne, France
{Christophe.Moulin, Cecile.Barat, Mathias.Gery, largeron, ducottet}@univ-st-etienne.fr

Abstract

This paper reports our multimedia information retrieval experiments carried out for the ImageCLEF track (ImageCLEFwiki). The task is to answer to user information needs, i.e. queries which may be composed of several modalities (text, image, concept) with ranked lists of relevant documents. The purpose of our experiments is twofold: firstly, our overall aim is to develop a multimedia document model combining text and/or image modalities. Secondly, we aim to compare results of our model using a multimedia query with a text only model.

Our multimedia document model is based on a vector of textual and visual terms. The textual terms correspond to words. The visual ones result from local colour descriptors which are automatically extracted and quantized by k-means, leading to an image vocabulary. They represent the colour property of an image region. To perform a query, we compute a similarity score between each document vector (textual + visual terms) and the query using the Okapi method based on the tf.idf approach.

We have submitted 6 runs either automatic or manual, using textual, visual or both information. Thanks to these 6 runs, we aim to study several aspects of our model, as the choice of the visual words and local features, the way of combining textual and visual words for a query and the performance improvements obtained when adding visual information to a pure textual model. Concerning the choice of the visual words, results show us that they are significant in some cases where the visualness of the query is meaningful. The conclusion about the combination of textual and visual words is surprising. We obtain worth results when we add directly the text to the visual words. Finally, results also inform that visual information bring complementary relevant documents that were not found with the text query. These initial results are promising and encourage the development of our multimedia model.

1 Introduction

The capacity of data storage increases constantly making possible the collection of large amount of information of all kinds, as texts, images, videos or combinations of them. In order to retrieve documents in such amount of data, information retrieval techniques tailored for the data types are required.

The ImageCLEF collection consists of multimedia documents made up of text and images. In this paper, we present our participation to the ImageCLEFwiki task. Our research goals are twofold: First, we aim to propose a multimedia document model combining text and image modalities adapted for multimedia retrieval. Second, we want to study the performance of our model

compared to a text retrieval approach. In order to benefit from our long time experience with textual model, we develop a vector-based model composed of textual and visual terms. The textual terms correspond to words of the text. The visual terms are obtained through a bag of words approach. Local colour descriptors are extracted from images and quantized by k-means leading to an image vocabulary.

After presenting our model, we will describe the runs we submitted. Then, we will comment on the results we obtained and conclude.

2 Visual and textual document model

ImageCLEFwiki is a multimedia collection where documents are composed of text and image. User needs are represented by queries ("topics"), which are also multimedia (text, image and concept). Hence a multimedia document model is necessary to handle such a collection. We focus our work on combining textual and visual information without using the concept field of the topics. Before explaining our model, we will describe the collection data.

2.1 Description of the data: ImageCLEF Wikipedia collection

The ImageCLEF Wikipedia collection is composed of 151'519 multimedia XML documents and 75 multimedia topics. The documents are made up of an image and a short text. Images are in a common format (jpeg and png) and their sizes are heterogeneous. They depict either drawings, paintings or screenshots. They are in colour or in black and white.
The textual part of a document is unstructured and consists of a description of the image, information about the Wikipedia user who has uploaded the image, or the copyright of the image. The average number of words per document is about 33 words.

It is also possible to get the whole Wikipedia article from which each image is extracted. The ImageCLEF Wikipedia collection provides 75 topics whose only 28 contains an image as example of the expected answer. For the text part, these queries were composed of about 2 or 3 words.

2.2 Textual representation model

One of the most known document model in textual information retrieval is the vector space model introduced by Salton and al. [2]. This model is based on a textual vocabulary $T = \{t_1, ..., t_j, ...t_{|T|}\}$. Each document is represented as a vector of weights $\vec{d_i} = (w_{i,1}, ..., w_{i,j}, ..., w_{i,|T|})$ where $w_{i,j}$ is the weight of the term t_j in the document d_i. In order to calculate the weight of a term t_j in a document d_i, a $tf.idf$ formula is usually applied. The $tf_{i,j}$ (term frequency) measures the relative frequency of a term t_j in a document d_i. We have used the one defined in the Okapi formula from Robertson and Jones [1]:

$$tf_{i,j} = \frac{(k_1 + 1) * n_{i,j}}{n_{i,j} + k_1 * (1 - b + b * \frac{|d_i|}{d_{avg}})}$$

where $k_1 = 1.2$ and $b = 0.75$ are two constants empirically defined in the Okapi formula, $n_{i,j}$ is the occurrence of the term t_j in the document d_i, $|d_i|$ is the size of the document d_i and d_{avg} is the average size of all documents in the corpus.
The idf_j (inverse document frequency) measures the discriminatory power of a term t_j and is defined as [1]:

$$idf_j = \log \frac{|D| - df_j + 0.5}{df_j + 0.5}$$

where $|D|$ is the size of the corpus and df_j is the number of documents in which the term t_j occurs at least one time.

The weight $w_{i,j}$ is then obtained by multiplying $tf_{i,j}$ and idf_j. This weight is higher when the term t_j is frequent in the document d_i but rare in the others. In our case, the size of our vocabulary T is 217'323 after applying a Porter stemming. The indexing has been performed with the Lemur software[1].

2.3 Visual representation model

In order to combine the visual information with the textual one, the images are also represented by a vector of visual words. It is therefore necessary to create a visual vocabulary $V\{v_1, ..., v_j, ..., v_{|T|}\}$. Our method consists in partitioning all images into 16x16 grids, a minimum of 8x8 pixels being required for each cell. It leads to about 256 cells per image, or about 38 million over all images.

For each cell, we compute a feature vector that contains the colour properties of the region. The vector is a 6 dimensional vector. The 6 dimensions correspond to the mean and the standard deviation for $\frac{R}{R+G+B}$, $\frac{G}{R+G+B}$ and $\frac{R+G+B}{3*255}$ where R, G and B are the red, green and blue components of the cell.

We apply a k-means algorithm over 4 millions of cells randomly selected within the 38 millions of cells to obtain 2'000 visual terms, which correspond to our visual vocabulary V. 2'000 has been chosen arbitrarily while 4 millions corresponded to the maximum number of cells we could compute due to the complexity of the k-means. Each visual term represents a cluster of feature vectors.

Then, each new image can be represented using a vector of visual terms. It is decomposed into a 16x16 grid and the local features are computed. Each cell is then assigned to the closest visual term from our visual vocabolary V, using the euclidean distance.

3 Experiments

Using the model described in the last section, we present our approach for multimedia documents retrieval from multimedia queries. Then we describe the runs we submitted to ImageCLEFwiki in order to evaluate our model.

3.1 Queries and matching

As mentionned before, the ImageCLEFwiki topics are composed of text, image and concept modalities. However, our model is designed to take only into account text and image modalities. Our retrieval approach consists in computing a similarity score between each document d_i and the query q_k using the Okapi method [1]. Documents are then ranked according to their scores. The following expression is used to compute the score:

$$score(q_k, d_i) = \sum_{u_j \in q_k} tf_{i,j} * idf_j * qtw_{k,j}$$

where $qtw_{k,j}$ is defined as:

$$qtw_{k,j} = \frac{(k_3 + 1) * n_{k,j}}{k_3 + n_{k,j}}$$

where $k_3 = 7$ is a constant defined in the okapi formula and where $n_{k,j}$ represents the occurrence of the term u_j in the query q_k. This score is higher when the term u_j is frequent in the document d_i but rare in the others and is weighted by the occurrence of u_j in the query.

Let us insist on the fact that the term u_j can be either a textual term t_j or a visual term v_j and that queries can be composed of textual terms only, visual terms only or both textual and

[1] `http://www.lemurproject.com`

visual terms which allows to perform text only queries, image only queries or multimedia queries.

The textual terms used for queries are those provided with topics. When visual terms are used, they are extracted either from the topic images or from the collection images. This will be detailed in the following section.

3.2 Submitted runs

We have submitted 6 runs to ImageCLEFwiki 2008, labelled from run 01 to run 06. Part of them are automatic (*auto*), others required manual selection of some relevant documents (*man*). Thanks to these 6 runs, we aim to study several aspects of our model, as the choice of the visual words and local features, the combination of textual and visual words for a query and the performance improvements obtained when adding visual information to a pure textual model. All these runs are summed up in table 1.

run name	first query (R_1)	run type	R_1use	second query (R_2)	results
LaHC_run01	t	*auto*	-	-	R_1
LaHC_run02	t	*auto*	v_{10}	v_t	R_2
LaHC_run03	t	*man*	v_{100}	v_t	R_2
LaHC_run04	t	*auto* *man*	- v_{100}	$t+ \begin{cases} v_q & \text{if } i_q \text{ exists} \\ v_t & \text{else} \end{cases}$	R_2
LaHC_run05	t	*man*	v_{100}	$t+v_t$	R_2
LaHC_run06	t	*auto*	v_{10}	v_t	$R_1 \cap R_2$

- t: text only query: $u_j \in q \cap T$
- R_1: first query results (baseline results); R_2: second query results
- *auto*: automatic run; *man*: selection of relevant images by user
- v_{10}: automatic selection of the first 10 results from R_1
- v_{100}: manual selection of the relevant documents in the first 100 results from R_1
- i_q: query image
- v_q: visual words extracted from the query image
- v_t: v_{10} or v_{100}

Table 1: Summary of the runs

We define a baseline that corresponds to a pure text model. It uses only textual terms for the query and scoring of documents. This run is run 01 and its results are noted R_1. We did not use neither feedback nor query expansion for this automatic run.

All other runs exploit both textual and visual information of documents. They consist in two successive queries. The first one corresponds to the baseline (R_1), while the second one is a visual or textual and visual query, either automatic or manual (R_2). R_2 is automatic when visual terms are extracted automatically from some images: we chose to select all the visual words of the top 10 retrieved documents issued from the baseline (v_{10}), assuming these results as relevant. R_2 is manual when the user is asked to select relevant documents among the first 100 results of $R_1(v_{100})$. There is no limit in the number of selected documents.

Run 02 and run 06 are automatic and only visual runs. For run 02, all the results of the second query are given as answer (R_2) while for run 06 an intersection is performed between the results of the first and second query ($R_1 \cap R_2$). This intersection is interesting as it emphasises the gain of the visual information use. Indeed, if the intersection is not null, it means that the visual query leads us to find documents which were not in the baseline results.

Run 03, run 04 and run 05 are manual runs. Run 03 is visual, while run 04 and run 05 are multimedia. For run 03, all the visual words of the selected images are used to perform the second query. For run 04 and run 05 we did a query expansion in order to analyse the combination of textual and visual information ($t+ v_{100}$). We kept the textual words of the initial query and added some visual words. For run 05, these words come from the manual selected images. The run 04 proceeds as run 05, unless a query image is provided for the considered topic. In that case, the second query is composed of the visual words extracted from the query image. Thanks to these two runs, we can study the influence of the number of relevant images used for queries.

4 Results

Table 2 shows that our best textual run (LaHC_run01) ranks us 6th on 12 participants. This run is ranked 22th on the whole 77 runs.

Rank	Participant	Run	Feedback/Expansion	MAP	P@10
1	upeking	zzhou3	QE	0.3444	0.4760
3	ualicante	IRnNoCamel	NOFB	0.2700	0.3893
4	cea	ceaTxt	QE	0.2632	0.4427
11	sztaki	bp_acad_textonly_qe	QE	0.2546	0.3720
13	cwi	cwi_lm_txt	NOFB	0.2528	0.3427
22	curien	LaHC_run01	NOFB	0.2453	0.3680
30	chemnitz	cut-txt-a	NOFB	0.2166	0.3440
44	imperial	SimpleText	NOFB	0.1918	0.3240
48	irit	SigRunText	NOFB	0.1652	0.2880
52	ugeneva	unige_text_baseline	NOFB	0.1440	0.2053
56	upmc-lip6	TFUSION_TFIDF_LM	NOFB	0.1193	0.2160
70	utoulon	LSIS_TXT_method1	NOFB	0.0399	0.0467

Table 2: Best textual run of each participant

As we can see on table 3, our best text+image run (LaHC_run03) ranks us 4th on 7 participants. This run, ranked 57th on the whole 77 runs, is outperformed by our best textual run (LaHC_run01).

Rank	Participant	Run	Feedback/Expansion	MAP	P@10
10	sztaki	bp_acad_avg5	NOFB	0.2551	0.3653
27	chemnitz	cut-mix-qe	QE	0.2195	0.3627
53	imperial	ImageText	NOFB	0.1225	0.2213
57	curien	LaHC_run03	FB	0.1174	0.2613
62	upmc-lip6	TIFUSION_LMTF_COS	NOFB	0.1050	0.2267
68	upeking	zhou_ynli_tliu1.1	FBQE	0.0603	0.0040
71	utoulon	LSIS4-TXTIMGAUTONOFB	NOFB	0.0296	0.0307

Table 3: Best text+image run of each participant

These results provide a basis to evaluate our approach on different aspects: the choice of the visual words, the way to combine textual and visual words and the performance improvements obtained when adding visual information to a pure textual model.

Concerning the choice of the visual words, results from run 02 and run 03 show us that they are meaningful in some particular cases. Indeed, using only the visual words, we are able to retrieve relevant documents which were not used for the query. For example, in run 03, we retrieved 42 relevant documents for the query "blue flowers" whereas we had just selected 9 images. However

there are some topics for which the results are bad. Only a third of the topics leads to more relevant documents than the number used for the query. This is due to the visualness of the query. It is obvious that for a query like "blue flower", the visual information is more useful than for a query like "peace anti-war protest".

Moreover, we observe from the results of run 04 and run 05 that only one image is insufficient to obtain good results. For topics which provide an image query, the results are always worth than the obtained results with several selected images. This can be explained by the fact that image query were not enough representative. Furthermore it is obvious that one image can not be enough expressive compared to several relevant images. The conclusion about the combination of textual and visual words is surprising. The comparison between run 03 and run 05 tells us that adding directly the text to the same visual words leads to worth results. The explanation is not obvious, but it seems to be that the proportion between textual and visual parts is unbalanced. Where we only have 2 or 3 textual words, we have several thousand of visual words.

Results also show that visual information bring complementary relevant documents that were not found with the text query. These initial results are promising and encourage the development of our multimedia model. The comparison between run 01, run 02 and run 06 inform us that 214 new relevant documents were retrieved with the visual query only over all topics. To take an example, 32 documents were found with run 01 and 30 with run 02. The intersection of both results coming from run 06 gives 13 shared documents. Thus, 17 relevant documents were retrieved with the visual information only.

5 Conclusion

We proposed a vector based model for multimedia documents. Thanks to the ImageCLEFwiki collection, we were able to test our model as this collection provide visual and textual information. We obtained encouraging results with visual words only based on coloured features. However, we did not take into account the specificity of the collection. For example, a lot of information as the copyright of the image is useless information for a retrieval task. Moreover, image names could be studied in order to improve search. For example, TheWhiteHouse.jpg could be replace by The White House.

For future work, as our local image features are basic, other features such as texture and edge information would surely lead to performance improvements. We also plan to automatically select the number of visual words using machine learning approaches. Finally, we aim to combine more efficiently textual and visual information.

6 Acknowledgements

This work is supported by the LIMA project[2] and the Web Intelligence project[3] which are 2 Rhône-Alpes region projects.

References

[1] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at trec-3. In *Text REtrieval Conference*, pages 21–30, 1994.

[2] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[2]LIMA project: `http://liris.cnrs.fr/lima/`
[3]WI project: `http://www.web-intelligence-rhone-alpes.org/`