Abstract

This paper presents a multimedia retrieval method which combines textual information with image features to describe the content of multimedia documents. Text and image are processed separately using a same bag-of-words approach and a tf.idf weighting scheme. We obtain two vectors of textual and visual terms that are linearly combined. We study the sensitivity of several factors which could impact on the performance of our model. We evaluate our method on the ImageCLEF’08 collection of about 150’000 Wikipedia documents. Experiments show that our model is competitive compared to other participants and yields top performance (2nd rank). Moreover, results prove that a multimedia model outperforms a text only one, which encourages to use multiple modalities rather than a single one.

Index Terms— Multimedia document retrieval, bag-of-visual words, tf.idf, vector space model, large collections

1. Introduction

In recent years, we observe a fast growth of large multimedia document collections, such as multimedia documents available through the World Wide Web. There is a need for an effective access to such collections. This requires the development of retrieval systems able to combine different modalities (text, image, video, etc.). Evaluation of the systems can be performed using benchmarks such as TREC, TRECvid, ImageEval or ImageCLEF which provide some evaluation frameworks for retrieval tasks.

Traditional systems exploit only the textual part of multimedia documents. A standard approach for text-based information retrieval is the weighted bag-of-words approach [1]. The challenge is now to find appropriate strategies to extend to other modalities, especially images since current available multimedia documents predominantly contain text and images. To retrieve images, there is a trend in using a similar bag-of-words approach to model images [2]. This technique has proven to be successful for image annotation [3] or scene classification [4].

In this paper, we aim to propose a multimedia document representation that combines textual and visual information for accurate multimedia document indexing and retrieval. Based on the bag-of-words approach and using very simple visual features, our system allows text and/or image queries and retrieves similar documents to user’s query. This work extends our ImageCLEF’08 paper [5]. The issues studied here focus on the feature parameter choice for the visual words, the visual vocabulary dimension and the visual-textual weighting and combination, which are crucial to the retrieval performance. Experiments are conducted on the ImageCLEF’08 multimedia document collection [6] to study the efficiency of the system and the influence of the previously mentioned parameters on the performance. An application of our approach to multimedia retrieval of images is described.

In Section 2, we detail our multimedia document model and the retrieval approach. In Section 3, we describe the test collection built to evaluate our approach and to study the influence of the different parameters of our model. In Section 4, we present some experiments and results on the ImageCLEF’08 retrieval task. Finally conclusions and future work are drawn in Section 5.

2. Model Overview

This section details our multimedia document model. Text and image are processed separately using a bag-of-words approach. They are then both represented as a vector of tf.idf weights characterizing the frequencies of each textual or visual terms. Using a same representation enables to combine the different modalities to perform multimedia queries and retrieval.

2.1. The TF.IDF weighting

As in the vector space model introduced by Salton et al. [1], we represent a document $d_i$ as a vector of weights $\tilde{d}_i = (w_{i,1}, ..., w_{i,j}, ..., w_{i,J})$. These weights characterize the importance of index terms for each documents. We explain firstly what the index terms are and secondly how the weights are computed. Index terms are derived from words
of the documents applying the Porter stemming to remove common word endings [7]. They constitute a vocabulary $T = \{t_1, ..., t_I, ..., t_T\}$. Indexing is then performed to represent documents with words of the vocabulary using the Lemur software [8].

The tf.idf term weighting scheme allows to measure the importance of terms. The more frequently a term appears in a document, the more important the term is to that document, while the term becomes less important when it often occurs in the collection. The importance of a term $t_j$ within the particular document $d_i$ is measured by the term frequency $tf_{i,j}$ while its importance over the corpus is evaluated with the inverse document frequency $idf_j$. The weight $w_{i,j}$ corresponds to the product of $tf_{i,j}$ by $idf_j$.

We use $tf_{i,j}$ and $idf_{j}$ defined in the Okapi formula by Robertson et al. [9] by:

$$tf_{i,j} = \frac{k_1 n_{i,j}}{n_{i,j} + k_2 (1 - b + b \frac{idf_j}{d_{avg}})}$$

where $n_{i,j}$ is the occurrence of the term $t_j$ in the document $d_i$, $|d_i|$ the size of the document, $d_{avg}$ the average size of all documents in the corpus and $k_1$, $k_2$ and $b$ are three constants.

$$idf_j = \log \frac{|D| - |\{d_i | t_j \in d_i\}| + 0.5}{|\{d_i | t_j \in d_i\}| + 0.5}$$

where $|D|$ is the size of the corpus and $|\{d_i | t_j \in d_i\}|$ the number of documents where the term $t_j$ occurs at least one time.

In the vector space model, by viewing a query as a short document, we can represent it as a vector of weights. A scoring mechanism enables then to compute a matching score between the query vector and a document vector. Scores are used to rank all documents for a given query $q_k$. We compare two scoring methods differing by the weights on the query words. One uses $tf$ weights ($score^1$) as in Robertson and al [9]. The other one uses $tf.idf$ weights ($score^2$) as in [8]. Detailed definitions and parameters values are summed up in table 1. Note that for $tf_{k,j}$, $b = 0$ because $|d_k|$ and $d_{avg}$ are not defined for a query.

### Table 1. Scoring equations and their parameters.

<table>
<thead>
<tr>
<th>Scoring</th>
<th>Parameters</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$score^1(q_k, d_i) = \sum_{t_j \in q_k} tf_{i,j} idf_j t_{f_{k,j}}$</td>
<td>$k_1 = 2.2$</td>
<td>$k_1 = 8$</td>
</tr>
<tr>
<td></td>
<td>$k_2 = 1.2$</td>
<td>$k_2 = 7$</td>
</tr>
<tr>
<td></td>
<td>$b = 0.75$</td>
<td>$b = 0$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scoring</th>
<th>Parameters</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>$score^2(q_k, d_i) = \sum_{t_j \in q_k} tf_{i,j} idf_j t_{f_{k,j}} idf_j$</td>
<td>$k_1 = 1$</td>
<td>$k_1 = 1$</td>
</tr>
<tr>
<td></td>
<td>$k_2 = 1$</td>
<td>$k_2 = 1$</td>
</tr>
<tr>
<td></td>
<td>$b = 0.5$</td>
<td>$b = 0$</td>
</tr>
</tbody>
</table>

2.2. Bag of visual words representation of images

The visual vocabulary $V = \{v_1, ..., v_j, ..., v_{|V|}\}$ is defined using a bag of words approach. First, each image is partitioned into a regular grid with $n_c \times n_r$ cells, $n_c$ and $n_r$ being the number of columns and rows. A minimum of $8 \times 8$ pixels is required for each cell. This uniform partitioning has the advantage of requiring low computational complexity. Next, local features are extracted from each cell to describe its color and texture properties. We choose to compute 6 features equal to the mean and the standard deviation for $R$, $G$ and $B$ and $R^2 + G^2 + B^2$, where $R$, $G$ and $B$ are the red, green and blue components of the cell. Then, we apply a k-means algorithm over all the computed cells to obtain $k$ clusters of features. Each cluster corresponds to what we refer to as a visual word. Each new image can be represented using a vector of visual terms. It is decomposed into a $n_c \times n_r$ grid and the local features are computed. Each cell is then assigned to the closest visual word using the euclidean distance. The image is finally represented by a vector of $tf.idf$ weight computed in the same way as text.

2.3. Combining text and image modalities

Using our two vocabularies $T$ and $V$, we compute the score as a linear combination of the scores corresponding to the two modalities. The $\alpha$ parameter lets us to add more or less visual information.

$$score(q_k, d_i) = \alpha \cdot score_T(q_k, d_i) + (1 - \alpha) \cdot score_V(q_k, d_i)$$

3. EVALUATION AND PARAMETER SETTING OF THE MODEL

The aim of this section is first to study the influence of the parameters of the previous model and second to evaluate the efficiency of the text-image combination. We also propose a parameter set which is optimal and adapted to the model. For that purpose, we use an evaluation collection composed of documents and queries for which we have a ground truth. This collection is made of a sub-set of ImageCLEF’08 one. The evaluation is performed calculating the mean average precision (MAP) which is a common criterion in information retrieval field.

3.1. Collections description

The ImageCLEF’08 collection is composed of 151’519 multimedia xml documents extracted from Wikipedia and 75 multimedia topics [6]. The documents are made up of an image and a short text. Images have heterogeneous sizes and depict either photos, drawings or screenshots. The textual part of a document is unstructured and consists of a description of
3.2. Influence of visual words parameters

Our visual vocabulary depends on two parameters: the number of cells (fixed by $n_c$ and $n_r$) and the number $k$ of clusters in the k-means clustering. In fact, the number of cells depends only on a single parameter since we choose to set $n_c = n_r$. Figure 1 presents the evolution of the MAP with respect to $k$ for different $n_c$ values. The scores are computed using $score^1$ formula.

Globally the MAP first increases quickly with $k$ then roughly increases slowly, reaches a maximum and roughly decreases slowly. For a given number of cells, we can determine an optimal value of $k$ ranging from about 1000 to 10000. As a function of the number of cells, the MAP value at the optimal $k$ varies. The best MAP is obtained for $n_c = 24$ and $k = 10000$.

![Fig. 1. MAP vs number of visual words at different cell sizes.](image)

3.3. Influence of the scoring method

The scoring formula of section 2.1 are originally defined for the textual information retrieval. It is interesting to know how they perform on visual information retrieval. For that purpose, we first compute for each of the two scoring formula, the optimal $n_c$ and $k$ parameters. As we can see on table 2, the MAP obtained with $score^2$ is better than the one of $score^1$. The better performance of $score^2$ seems due the use of a $tf.idf$ weighting in the query.

<table>
<thead>
<tr>
<th>scoring</th>
<th>$n_c$, $n_r$</th>
<th>$k$</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$score^1$</td>
<td>24</td>
<td>10000</td>
<td>0.0135</td>
</tr>
<tr>
<td>$score^2$</td>
<td>16</td>
<td>10000</td>
<td>0.0163</td>
</tr>
</tbody>
</table>

Table 2. Optimal MAP and visual words parameters depending on the scoring.

3.4. Influence of the fusion parameter and conclusion

The fusion between textual and visual information depends on the $\alpha$ parameter defined previously. Using the optimal parameters for $n_c$, $n_r$ (16) and $k$ (10000) and using the best scoring formula ($score^2$) for the two modalities, we compute the MAP with respect to $\alpha$ parameter. We can see on figure 2 that the MAP reaches a maximum for $\alpha = 0.015$.

These results allow us to conclude on the efficiency of the approach: adding a few visual information to the textual one improves the retrieval system.

![Fig. 2. MAP vs $\alpha$ for optimal text/image combination.](image)

4. EVALUATION ON AN INFORMATION RETRIEVAL TASK

Using the optimal parameters computed on the subcollection, we evaluate our model on the multimedia information retrieval task of ImageCLEF’08. Then, we compare these new results with those from the competition [5].

4.1. Overall results analysis

In the original collection, topic images were not always provided. In order to use the visual information, we choose the first two relevant images among results returned by the textual queries. We use the corresponding visual words as visual query. It would correspond to a relevance feedback approach where the user selects the first two relevant images in order to enhance results. Figure 3 shows obtained results for the text.
only query, the visual only query and the fusion. These precision/recall curves show that the combination of the textual and the visual results leads to a significant improvement.

Fig. 3. Precision/recall curve for different modalities.

4.2. Results analysis over the competition

When we participated to the ImageCLEF’08 competition [6], we did not use the fusion approach presented in this paper and our best result used text only queries. To evaluate our model, we compare the new results with our previous ones and with those of other participants. Table 3 shows that our new results obtain the second rank between upeking and cea. One can note that their runs were only based on the use of textual information. They performed query expansion using extra data as thesaurus not provided by the competition. This demonstrates that using visual information is a very promising approach.

Table 3. Some ImageCLEF’08 results.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Participant</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>upeking</td>
<td>0.3444</td>
</tr>
<tr>
<td></td>
<td>fusion (α = 0.015)</td>
<td>0.3056</td>
</tr>
<tr>
<td>2</td>
<td>cea</td>
<td>0.2735</td>
</tr>
<tr>
<td>22</td>
<td>curien</td>
<td>0.2453</td>
</tr>
</tbody>
</table>

5. CONCLUSION

We proposed a vector based model for multimedia information retrieval adapted for documents composed of both text and image parts. This model was based on the tf.idf approach, classically used for text documents, applied both on the textual and the visual part. For the visual part, an image vocabulary was obtained through the clustering of very simple color features computed on a regular grid. The combination of the two modalities was achieved by a linear combination of the visual and textual scores. The model was evaluated and then applied on ImageCLEF’08 data set. The evaluation allowed to determine an optimal parameter set and to emphasize the improvement provided by the combination of textual and visual information. The results obtained on ImageCLEF’08 data set were compared to the results of the competition and showed that our system ranked at the second position.

For future work, we plan to combine other visual features as texture and shape parameters. Another perspective is to determine an optimal combination parameter for each query.

6. ACKNOWLEDGEMENTS

This work was partly supported by the LIMA project and the Web Intelligence project which are 2 Rhône-Alpes region projects of ISLE cluster.

7. REFERENCES