LIG at ImageCLEFphoto 2008

Philippe Mulhem
UJF, UMR CNRS 5217, Laboratoire d'informatique de Grenoble
Philippe.Mulhem@imag.fr

Abstract

This working notes describe the runs and results obtained by the LIG at ImageCLEFphoto 2008. The submitted runs are: two runs (text only and text+image) without diversification on classes, and two runs (text only and text+image) with class diversification were submitted. The text retrieval is based on language model of Information Retrieval, and the image part is processed using RGB histograms on 9 image blocks with a similarity value based on Jeffrey divergence. Results using text+image are obtained by a linear combination of normalized results on text and image. The diversification is based on clusters, according to the cluster given in the queries. When the cluster name is not directly extracted from the images (like city or country), we apply a visual clustering. Not surprisingly, the cluster recall at 20 (i.e., cr(20)) results are higher for the runs that include diversification. On the other hand, the precision at 20 and the mean average precision results are higher without diversification on our runs, for both text only and image+text 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

Text indexing and retrieval, Image indexing and retrieval

1 Introduction

This paper describes the runs that where submitted at ImageCLEFphoto 2008 by the LIG (Laboratoire d'Informatique de Grenoble). The runs submitted deal with text only and text+image retrieval. The main idea behind our submissions is to make a first step into considering language models for text and for image. Another aspect of this work is to check the impact of image content on the quality of the retrieval. When considering clusters of images to ensure diversification of the results, we applied simple solutions based on the analysis of the images description when available, and we relied on visual clustering of the images when the clustering was not expressed in the location field of the image description. Among the four runs submitted, our best results are below the average ImageCLEFphoto results for the map (-5%) and the precision at 20 documents (-3%), and the results obtained for the clustering recall at 20 documents are above the average

(+19%). This fact shows that the diversification based on simple features may impact positively the results.

The remaining of the paper is organized as follows. Section 2 describes the runs submitted by focusing on the text processing, the image processing and the fusion and the processing of text+image, section 3 details the results obtained, and we conclude in part 4.

2 Description of the runs

As descibed above, the runs submitted consider on ons sidetext only and text+image indexing and retrieval, and the use of diversification or not on other side. We describe in this section these different aspects.

2.1 Text processing

When considering text retrieval, it has been shown by [1] that language models (LM) of Information Retrieval, inspired by speech recognition, give results that are close to, or outperform, existing approaches (like Vector Space Model [3] or probabilistic models based on BM25 for instance [2]). Different kinds of Language Models already exist, and we chose to use a language model that expresses the probability for the query Q to be generated from a document model D by : P(Q|D). Such probability is computed using the probability of any term t_i to be generated by the document, P(t_i|D). One strong aspect of the LM is the use of smoothing leading to consider that a document that does not contain a term does not have a zero probability of generating this term. The smooting used in our experiments is the Dirichlet smoothing, that has been shown in citezhailaferty04 to be a good smothing method. Such smoothing is described using the following formula:

$$P(t_i|D) = \frac{\mathrm{tf}(t_i,D)+\mu.\frac{\mathrm{tf}(t_i,C)}{\mathrm{tf}(*,C)}}{\mathrm{tf}(t_i,D)+\mu}$$

where

- C is the corpus of documents considered,
- tf(x,D) denotes the term frequency of the term x in the document D,
- tf(x,D) denotes the term frequency of the term x in the corpus C,
- μ is a parameter defined experimentally that indicates the importance of the corpus on the smoothing.

A usual information retrieval preprocessing is performed: we remove stopwords from the image descriptions and from the queries, and we stem the texts using a Porter stemmer. We use the concatenation of the `TITLE`, `DESCRIPTION` and `NOTES` fields, in lower case, of the xml description of the images as their text description.

For the query processing, we used the query and the narrative as the text query, and the smae preprocessing is applied on this text.

As explained above, the query processing evaluates P(Q|D) as:

$$\begin{aligned} P(Q|D) &= P(t_{q,1}, ..., t_{q,nq}|D) \\ &= \prod_i P(t_{q,i}|D) \end{aligned} \tag{1}$$

The results are ranked according to the decreasing order of this probability.

2.2 Image processing

The processing of the images computes histograms on image blocks. The result obtained on 9 blocks gave better results than taking the whole image in previous of our studies on personal photographs.

Each image of the corpus is split into a 3x3 regular grid. For each of the blocks $b_{I,i}, 1 \leq i \leq 9$ of an image I, we extract an RGB histogram, $H_{I,i}, 1 \leq i \leq 9$. The choice of the RGB color space is due to the fact that it gives similar results than other color spaces that require more processing time to be computed. Each histogram has 512 bins, according to a 8×8×8 regular split of the RGB cube. Such size is a tradeof between too small histograms unable to discriminate visual elements, storage uasge and matching duration. Then, one global histogram GH_I with 512×9=4602 bins is created for one image, by concatenating the 9 $H_{I,i}$ histograms. GH_I is then normalized to 1.

The matching function between two images I and J uses the Jensen-Shannon divergence. This divergence is a symmetrical version of the Kullback-Leibler divergence. The formula of the Jensen-Shannon divergence between two histograms GH_I and GH_J is:

$$JS(GH_I||GH_J) = \frac{1}{2}\sum_i GH_I(i).log(\frac{GH_I(i)}{\frac{GH_I(i)+GH_J(i)}{2}}) + \frac{1}{2}\sum_i GH_J(i).log(\frac{GH_J(i)}{\frac{GH_I(i)+GH_J(i)}{2}}) \quad (2)$$

The similarity $S_{im}(I,J)$ between two images I and J is defined as $1 - JS(GH_I||GH_J)$.

In imageCLEFphoto, the visual part of a query is composed of 3 images, each image representing one sample of what is expexted. We consider that the relevance of images in the corpus depends on the best relevance value wrt. each of the query images. The similarity between a set of images $IS_q = \{I_{q,i}\}$ and one image J is then defined as the maximum of the similarity between J and each of the images of IS_q.

The results are then ranked according to the decreasing order of this similarity.

2.3 Image+Text processing

Several ways may be used to represent mixed textual and visual data for retrieval. We consider that even if the internal representation for these two media are somewhat similar (distribution of probabilities), their inner nature refrain us to apply a kind of early fusion of them by concatenating the two distributions. Another reason is that "adding" two distributions does not necessarily create another one. That is why, in our run, we consider a late fusion defined as a linear combination of the two results (text and image) obtained.

In fact, the linear distribution is applied on normalized results, using for the text (resp. the image) the minimum and the maximum matching values to ensure the normalized results to be in [0, 1]. This linear combination is refined using the following heuristics: we assume that the best visual results (i.e., very similar to one of the query images), are often relevant, which is not the case for the text. So, we extend the matching function using a condition on the matching:

$$matching(Q,D) = \begin{cases} 1, & \text{if } Sim_v(Q_v,D_v) > t_v \\ \alpha.Sim_t(Q_t,D_t) + (1-\alpha).Sim_v(Q_v,D_v) & \text{othewise} \end{cases} \quad (3)$$

with Q_v (resp. Q_t) the visual (resp. textal) part of the query, and D_v (resp. D_t) the visual (resp. textal) part of the document.

The results are then ranked according to the decreasing order of this similarity.

2.4 Diversification processing

As one task for the imageCLEFphoto this year was dedicated to study the capacity of the systems to provide diverse results for a query, and not only near duplicate results. To achieve this goal, we defined clusters according to several criteria:

- country: based on the LOCATION field of the corpus images description, we generated a set of clusters $C_{country}$, that groups images by country name. $C_{country}$ contains 23 classes, with on average 952 images per cluster.
- city: based on the LOCATION field of the corpus images description, we generated a set of clusters C_{city}, that groups images by city name and country when available. One cluster of C_{city} contains all the images with no city. C_{city} contains 511 classes, with on average 38 images per cluster.
- clusters on date where also built. Because such clusters were not used during the query processing we only mention them;
- visual: we used a KNN clustering, namely C_{visual}, based on the visual description of the images (4608 dimension histograms). The target number of cluster is 500, i.e. similar to the number of clusters for the cities. There is 40 images per cluster on average.

Depending on the query un ser consideration, a clustering is chosen for the diversification process:

- Queries having a cluster name `city` are diversified using the C_{city} clusters;
- Queries having a cluster name `state` or `country` are diversified using the $C_{country}$ clusters;
- Queries having another cluster name are diversified using visual clusters C_{visual}.

We describe now the way the diversification is processed, according to a clustering C_x:

1. From the top of the initial (not diversified) result list L_o, we generate a diversified list L_d that contains only the best representative of one cluster of C_x according to the ranking of L_o, until the diversified list contains 20 elements.
2. In a second step, we loop on L_o from its top, and we add to L_d the elements of L_o that do not already belong to L_d.

This process ensures that the 20 first results are from different clusters.

3 Results

We describe in this section the official results obtained by our approach. According to the parameters described in the previous parts, the λ of the text language model is set to 1500, using the Zettair system [4]. For the runs text+image, we set the threshold t_v to 0.99 and the α used in the linear combination to 0.55 . We apply the diversification process explained earlier. To be able to study the advantages or drawbacks of the diversification, we submitted diversified and not diversified runs.

The figure 1 presents the evolution of the precision for our results, and the figure 2 presents the cluster recall for our four submitted runs. The runs without diversification (NOCLUST) are presented with dotted lines in these two figures, where the plain lines correspond to diversified runs.

We analyse first the results regarding the precision of figure 1. Without diversification, we see in this figure that the text+image configuration outperforms the text only runs (+34% averaged on the 6 values given by the official results).
For the text only run, we see that for the precision after 15, 20 and 30 documents the precision value is roughly constant (0.2171, 0.2026 and 0.2103). This means that at 15 documents the systems gives 3.2 relevant documents, and at 30 documents the system gives 6.3 relevant documents on average. On the other hand, the text+image run shows a monotonic decreasing of the precision. Considering the diversified results, the text+image also outperfoms the text only based system by 46.32% (averaged on the 6 values given by the official results). For both text+image and text

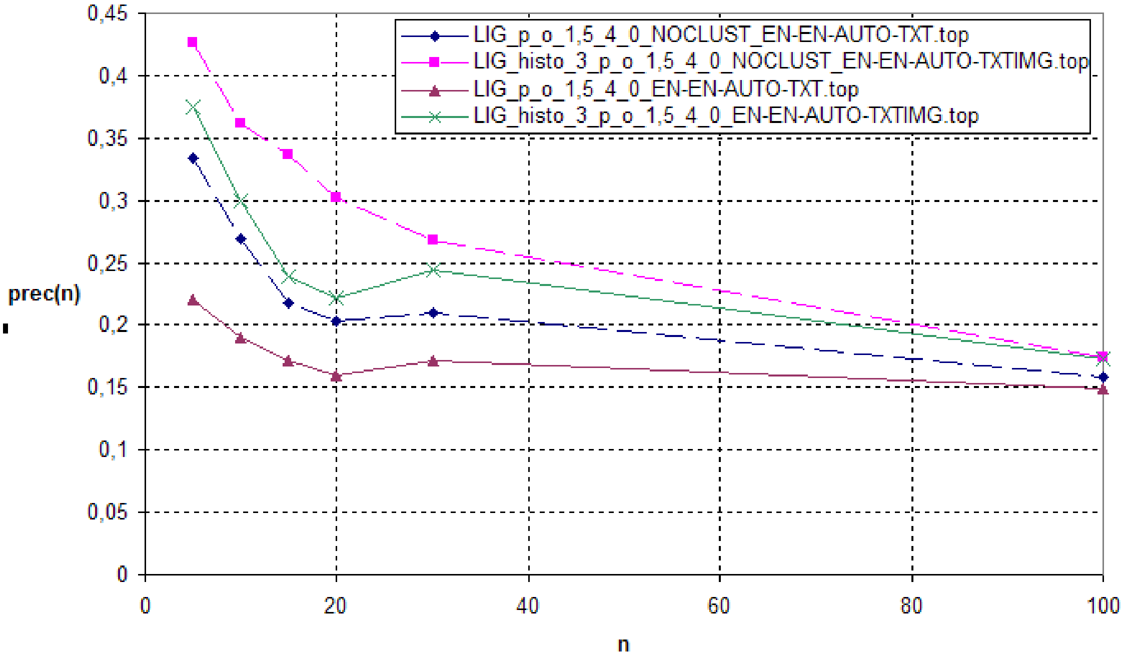

Figure 1: Precision for LIG runs for text only (TXT) and text+image (TXTIMG), without and with diversification

only results, we see that the precision is increased between 20 and 30 documents. We explain this fact by our diversification process that manages in a different way the first 20 documents, and then only adds the retrieved documents: because the first documents have more chances to be relevant, such documents can be placed from the 20th document of the diversified list, increasing the precision.
The diversification process lowers the precision values obtained, for both the text+image and the text only results. This can be explained by the fact that the diversification pushes up documents that are not in the first results of the initial list, and these documents are potentially less relevant. For the text+image results however, the difference after 30 documents is not large (0.2684 versus 0.2436), which means that on average only one additional relevant document is retrieved without diversification compared to the results after diversification.

We study then the cluster recall resuts presented in figure 2.
As expected, the results obtained after diversification for text only (resp. text+image) outperform the results obtained without diversification (+9.6% for text only retrieval, +5.7% for text+image, averaged on the 8 values of the official results). At 20 documents, the difference is +16.7% for the text only, and +10.8% for the text+image run. We see on the figure2 that, here also, the text+image runs give better results than text only runs. However, after 50 documents, the difference beween diversified and undiversified results is very small for the text+image runs: +2.0% on average. In this case, the diversification is almost useless with respect to the cluster recall. For the text only runs, a difference of 3.0% is achieved after 100 documents, leading to conclude that the undiversified text results benefit more from the diversification than the texte+image results. This is validated through the study of the average of the differences for diversified results versus undiversified results up to the 30th document (i.e., for 5, 10, 15, 20 and 30): the text only diversified run outperforms by +19.9% the undiversified text only, and the text+image diversified run outperforms by +11.2% the text+image undiversified results.

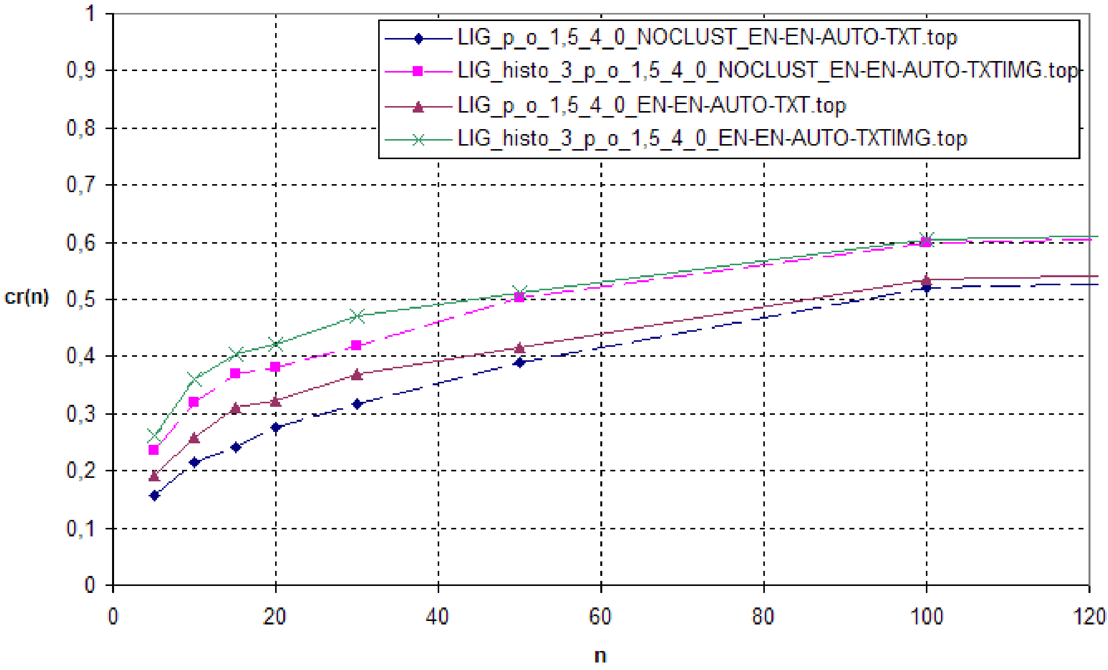

Figure 2: Cluster Recall for LIG runs for text only (TXT) and text+image (TXTIMG), without and with diversification

4 Conclusion

The work described here lists the runs and results obtained by the LIG at ImageCLEFphoto 2008. We experimented text and text+image retrieval. For the text, we considered a language model using a Dirichlet smooting. For the image, we extracted RGB histograms of 9 blocks of the images, ane the matching was performed using a Jeffrey divergence. For the retrieval and text+image, we used a linear combination of the matching values for the text and for the images of the query. We applied a simple diversification scheme in a way to force multiple clusters to be given for the 20 first results. We submitted results with and without diversification.

When comparing our different runs, we found out that text+image runs consistently outperform text only runs. For the precision values, undiversified results outeperfom diversified runs. For the cluster recall, diversified runs outperfom undiversified results. So, we conclude that precision and recall do not benefit both from the diversification process. Such antagonism between recall and precision is new in IR, but we think that studying the effect of our diversification process, and testing other ways to diversify (other clusters, smarter clustering, smarter reranking of the results) may help to increase both precision and recall, at least below 30 documents.

The results obtained are below the average of the runs with and without diversification for the text only, according to the precision and the cluster recall. For the text+image runs, the results are below the average when considering the non-diversified run, but above the average when using diversification for the cluster recall.

Acknowledgement

This work was partially supported by the French National Agency of Research (ANR-06-MDCA-002).

References

[1] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the ACM SIGIR*, pages 275–281, 1998.

[2] S E Robertson, S Walker, S Jones, M M Hancock-beaulieu, and M Gatford. Okapi at trec-3. pages 109–126, 1995.

[3] Gerard Salton and M.J. McGill. *Introduction to modern information retrieval.* McGraw-Hill, 1983.

[4] The Zettair search engine. http://www.seg.rmit.edu.au/zettair/.