Evaluation of Diversity-Focused Strategies for Multimedia Retrieval

Julien Ah-Pine, Gabriela Csurka, Jean-Michel Renders

► To cite this version:

Julien Ah-Pine, Gabriela Csurka, Jean-Michel Renders. Evaluation of Diversity-Focused Strategies for Multimedia Retrieval. Deselaers, Th., Ferro, N., Gonzalo, J., Kurimo, M., Mandl, Th., Petras, V. Evaluating systems for multilingual and multimodal information access, 5706, Springer, pp.677-684, 2009, Lecture Notes in Computer Science, 978-3-642-04447-2. hal-01504452

HAL Id: hal-01504452

<https://hal.science/hal-01504452>

Submitted on 27 Apr 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Evaluation of Diversity-focused Strategies for Multimedia Retrieval

Julien Ah-Pine, Gabriela Csurka and Jean-Michel Renders

Xerox Research Centre Europe
6 ch. de Maupertuis
38240 Meylan, France
`FirstName.LastName@xrce.xerox.com`

Abstract. In this paper, we propose and evaluate different strategies to promote diversity in the top results of multimedia retrieval systems. These strategies consist in clustering, explicitly or implicitly, the elements of the top list of some initial ranking and produce a re-ranking that favours elements belonging to different clusters. We evaluate these strategies in the particular case of ImageCLEFPhoto 2008 Collection. Results show that most of these strategies succeed in increasing a diversity performance measure, while keeping or slightly degrading precision of the top list and, more interestingly, they achieve this in complementary ways.

1 Introduction

Up to now, many multi-modal retrieval systems have focused on addressing the semantic gap between textual and visual contents of multimedia documents. The scientific challenges are traditionally oriented towards exploiting in an efficient manner these two types of information in order to improve the search results. In ImageCLEFPhoto 2008, however, the main goal is to promote diversity among the first search results. We address this issue using a two-step approach. In the first step, we ignore the question of diversity. In other words, we first try to find the most relevant objects using the material introduced in [1]. Then, in a second step, we re-rank the first relevant objects by taking into account their mutual similarities in order to avoid redundancy and thus to promote diversity. These different approaches are described in details in [2]. Here, we will just briefly remind them, as the main goal of this paper is to perform a deep analysis of the results, as well as to compare and to discuss the strengths and weaknesses of the proposed approaches.

2 Incorporating Diversity in Retrieval Ranking

As mentioned previously, one of the aims of ImageCLEFPhoto 2008 compared to previous sessions, is to promote diversity in the search results so that the first retrieved elements are not redundant. Nevertheless, let us begin by recalling our basic multimedia retrieval strategy, that does not take into account the

redundancy in the returned ranked list. Complete description and rationale can be found in [2]. Remind that diversity-based extensions will be built on top of this method.

As textual similarity measures, we used the cross-entropy between the language models of two objects [3], the documents of the collection being expanded by thesaurus enrichment¹, while the query was expanded by pseudo-relevance feedback. As visual similarity measures, a cosine-like norm between the “Fisher Vector” representation of two images was employed [4]. Based on these mono-modal similarities, we developed an intermediate fusion method which efficiently combine them in the context of multimedia retrieval. This kind of fusion operator can be understood as query score regularization through a two-step diffusion process, the first step being performed in one mode and the second step being performed in the other one [1]. The ranking obtained with the basis method will be denoted by *xrce_cm_best_basic*.

Let us go back to the objective of promoting diversity in the first elements of the ranked retrieval list. To this end, we investigate two main families of methods: implicit and explicit clustering-based approaches. Both approaches use a pair-wise similarity matrix (designated by *Sim*) between objects of the collection in order to model the diversity. In ImageCLEFPhoto 2008, we tested two kinds of similarities. The first one, designated by *tilo* is purely textual and is computed by the cross-entropy measures between pairs of smoothed language models of the documents restricted to their *title* and *location* fields. The second one, denoted by *cm*, is a pair-wise cross-media similarity matrix that takes into account both textual and visual content in the same way that we did for computing the similarity between a query and an element of the collection.

The first method to avoid redundancy is commonly known as “Maximal Margin Relevance” (MMR) [5]. It amounts to re-rank the search results so that the element chosen at rank j has to be dissimilar to elements that were already selected at ranks $j' < j$. More concretely, given an initial relevance score vector $Score(q)$ (for a given query q , this is typically our *xrce_cm_best_basic* ranking), as well as a pair-wise similarity matrix *Sim*, the MMR framework supposes that the elements should be ranked by greedily choosing at each step (rank) j the element o^i that maximizes the following re-ranking criterion:

$$MMR(o^i) = \beta(j)Score(q, o^i) - (1 - \beta(j)) \max_{o^{i'} \in P_j} Sim(o^i, o^{i'}) \quad (1)$$

where $\beta(j)$ is a mixture parameter² (between 0 and 1) depending on the rank and P_j is the set of objects already selected (ranks lower than j).

¹ Using the English Open Office thesaurus available on <http://wiki.services.openoffice.org/wiki/Dictionaries>, giving 15 times more weights to the original words.

² Traditionally, β is kept constant, but we propose here a more efficient variant, where $\beta(j)$ linearly increases between $\beta(1) = \alpha (< 1)$ and $\beta(k)=1$ for some k (typically $k=100$), before saturating at value $\beta = 1$.

The second family of methods are based on an explicit clustering of the first k elements, followed by a strategy designed to re-rank the elements so that many different clusters are represented among the first elements of the re-ranked list. In this method, for a topic q , we start by selecting the first k elements according to $Score(q)$, which we denote by P_k . Next, we cluster objects within P_k in order to find different themes, using the *Sim* pair-wise similarity matrix as basic material. This is realized using the Relational Analysis approach [6, 7], which has particular advantages in this framework because it does not impose to fix a priori the number of clusters and because it can deal with strongly unbalanced cluster size distributions (by isolating very small, but significant clusters).

After having clustered the elements within P_k , the top elements of the initial list are re-ranked by the following strategy: we go top down through the elements of this list and, if the cluster id of the current object is not yet represented by other objects previously considered, it is put in the priority list; otherwise, it is put in the non-priority list; we iterate this operation until the number of different clusters represented in the priority list reaches *nbdiv*; after this, we naturally build a new list by taking the priority list, followed by the non-priority list (in each list, objects are ranked following their original relevance scores).

As far as nomenclature is concerned, to designate our ranking methods, the *tilo* suffix indicates that we used the purely textual (title + location) similarity measure to cluster the objects of the collection. The *cm* suffix refers to the use of the pair-wise cross-media similarity measure. Implicit and explicit clustering are differentiated by the *mmr* and *nbdiv* suffixes respectively; in the latter case, the maximum number of clusters represented in the top list is indicated just after.

3 Analysis of Experimental Results

3.1 Analysis of Retrieval Results

For details about the task and the collection, we refer to the overview paper [8]. Before analysing the effects of the diversity-focused strategies, it is interesting to analyse the basic ranking algorithm (*xrce_cm_best_basic*), in order to understand the strengths and weaknesses of the underlying method. That is the goal of this sub-section. Clearly, this algorithm outperforms very significantly the mono-modal strategies, as well as the late fusion approach (see our results in [1]). Let us have a deeper look on results query by query (see Figure 5).

First, we recall (see also [2]) that the only linguistic analyses made on the text or the query were lemmatization and stop-word removal.

Even if the lemmatization was in most cases beneficial, (e.g. for topic³ 55/35 Peru and Peruvian), we also noticed a negative effect in the case of the topics 39/24 and 40/25 where “*bad weather*” and the adjective “*weathered*” were considered as identical words after lemmatization. For this reason, texts containing words like “rain, hail, fog, wind, storm, etc” (making reference to bad weather,

³ We will refer to topics by t/n , where t refers to the topic number provided and n to its order in the topic list. In the figures the topics’ orders appear from 1 to 39.

Fig. 1. We can obviously see the visual similarity (apart weather conditions) between the query images of topic 39/24 (first three images) and three of the four top images retrieved by our system. The missing top retrieved image contained was one of the top ranked ‘weathered sandstone’ images.

Fig. 2. Top 10 images for the topic topic 2/1 with the `xrce_cm_best_basic` (top row) and `xrce_cm_mmr_07` (bottom row).

but not using explicitly the word “weather”) had much lower relevance scores than images containing “weathered sandstone”, even if the word “weather” was added by thesaurus enrichment (but with a low weight by construction). Moreover, from a purely visual viewpoint, “*bad weather*” is a concept rather difficult to be detected at least with our visual similarity measures as we can see in Figure 1. Indeed, top images look very similar to the query images apart from some small umbrellas that could be interpreted as a sign of bad weather.

The lack of semantic or linguistic analyses of queries made some of the queries quite difficult to be handled by our approach. We can mention here the topic 2/1, “*church with more than two towers*” where the notion of “more than two” would require a semantic analysis and a possible transformation to “three”, “four”, etc. Furthermore, visually, even if we retrieve the same building as present in a given query (e.g. the St Patrick Cathedral in Melbourne as shown in Figure 2⁴ from different points of view), it is difficult to ensure that all towers are visible in the image (the assessor appeared to rely more on the image than on her prior knowledge about the building). For buildings that are not in the query images, the task becomes even more difficult.

Similarly, for topic 18/12 entitled “*sport stadium outside Australia*”, as we removed “outside” as a stopword and did no semantic analysis of the query, texts mentioning “Australian sport stadiums” get higher textual ranks than any “sport stadiums” from other countries.

The failures in the above examples were rather due to the lack of semantic analysis or logical reasoning about the query content. Another cause of failure

⁴ In all figures, a colored box around an image means that the image is relevant and different colors make reference to different ground truth clusters.

Fig. 3. The first line represents the top 10 results provided by the `xrce_cm_best_basic` method for the query 20/14. The second line shows the top 10 results of the diversity-focused approach `xrce_cm_mmr_07`. Compared to the former, the latter was able to bring a relevant image to the 5th position from 70th (in the first case).

Fig. 4. The method `xrce_tilo_nbdiv_10` introduced diversity for the query 6/4 within the top 10 results (second line) compared to the `xrce_cm_best_basic` method keeping the same P10 performance.

stems from the cross-media pseudo-relevance feedback mechanism that is the basis for our cross-media similarity measure. The real limitation (and therefore risk) of the proposed cross-media similarity can be seen through the analysis of the topic 20/14, “*close-up of animals*”. As shown in Figure 3, the results were really poor and suprising. Trying to understand why, we realized that due to high visual similarity between the “golden eagle” close-up query image and some close-up “mummies” images (top 2), the textual query was enriched by the words related to “Nazca grave’s mummies” images. This resulted in pushing at the top similar “mummies” images. This was possible because there were no stronger candidates, as the original query virtually shared no textual content (for most animal images, the word “animal” does not appear explicitly) and no visual content with the documents of the collection.

Fortunately, this was a rare case, and in most cases the system was able to take advantage either from textual or from visual similarities (or both) and to successfully combine them. We can mention the excellent results obtained for queries 6/4 or 44/28 (see Figure 4) where the top results show alternatively documents that are visually similar or textually similar.

3.2 Analysis of Diversity Results

In this sub-section we focus our analysis on the different strategies we proposed for avoiding redundancy in the top lists. These comments are made on the basis of the Figure 5 given previously and the Table 1. The latter represents some of the best results we obtained for ImageCLEFPhoto 2008 (incidentally, they were also the best ones when compared to the other participants). Particularly,

Fig. 5. P20 and CR20 results per query for *xrce_cm_best_basic* (blue), *xrce_cm_mmr_07* (green) and *xrce_tilo_nbdiv_10* (red).

we will pay more attention to the methods denoted by *xrce_tilo_nbdiv_10* and *xrce_cm_mmr_07* that are respectively the best results we obtained with explicit and implicit clustering approaches.

The analysis we give below takes into account the P20 and CR20 measures, following the criteria that were chosen to evaluate the results during Image-CLEFPhoto 2008.

When seeking to promote diversity departing from the basic run *xrce_cm_best_basic*, we can observe in Table 1, that any diversity-focused method fails to increase, on average, the P20 measure. However, all methods perform better than the basic run regarding the CR20 measure. In other words, by trying to eliminate redundancy among the first retrieved objects, unfortunately, we might push relevant objects out of the 20 first re-ranked elements and on the contrary, we might put into this final top list some irrelevant objects. Nevertheless, this is an average trend. Indeed, according to Figure 5, it happens, that seeking for diversity also leads to improved precision.

Next, if we compare the results given by the two kinds of method (implicit and explicit clustering), we can notice, according to Table 1, that the former

Table 1. Some Best results among XRCE’s runs

Rank	Run	P10	CR10	P20	CR20	P100	CR100
1	xrce_tilo_nbdiv_10	0.5103	0.3636	0.5423	0.4146	0.2831	0.6374
2	xrce_cm_nbdiv_10	0.4923	0.3608	0.5474	0.4111	0.2831	0.6374
4	xrce_cm_mmr_07	0.6538	0.3012	0.5500	0.4015	0.2790	0.6407
7	xrce_tilo_mmr_07	0.6462	0.3076	0.5385	0.4006	0.2751	0.6492
12	xrce_cm_best_basic	0.6846	0.2816	0.5731	0.3727	0.2831	0.6374

Fig. 6. The method `xrce_tilo_nbdiv_10` introduced diversity for the query 3/2 within the top 10 results (second line) compared to the `xrce_cm_best_basic` method but decreasing the P10 performance.

has, on average, a better P20 than the latter. On the contrary, with respect to CR20, explicit clustering method generally performs better.

Then, let us look at Figure 5 more closely. The explicit clustering strategy exhibits a consistent, stable behaviour, where it systematically gives slightly lower or equal P20 performance than the basic ranking, while offering CR20 performance that are superior or equal to the baseline. In that context, Figure 4 that illustrates the results of topic 11/6 is a good example that demonstrates the benefits of the explicit clustering method for avoiding redundancy.

Implicit clustering does not offer such a stability in its behaviour. In fact, the implicit clustering method seems to “take more risk” in the re-ranking process with a diversity seeking goal than the explicit clustering method, with a consequence of increased variance in the performance. In that context, the topic 20/14 illustrated in Figure 3, is a good example that shows the benefits of the implicit clustering. Indeed, in that case neither the basic run nor the clustering explicit runs allow to find a single relevant object in their top lists. Despite this fact, the implicit clustering had allowed to find a relevant objects in its top 10.

Finally, we illustrate in Figure 6 another issue, related to the fact that the clustering process can result in a totally different thematic than the one intended by the organizers. For the topic 3/2, entitled “*religious statue in the foreground*”, the ranking `xrce_tilo_nbdiv_10` (based on purely textual information) offers diversity (in terms of location), but different from what was expected (types of religious statues).

4 Conclusion

From a pure relevance-based assessment, our method based on cross-media relevance feedback was able to adequately combine and leverage textual and visual information in the query and the documents. Incapacity to give good results was in general due to the lack of deeper logical and semantic analysis of the query content. When considering diversity-focused strategies, three main lessons can be learned. First of all, these strategies are very dependent on the results of the basic ranking (they often fail to capture relevant elements that are low in the list). Secondly, implicit and explicit clustering strategies are often complementary (they are not beneficial for the same type of queries). Finally, implicit clustering strategies favour more exploration than explicit clustering strategies, with the consequence of exhibiting larger variance in the performance results. Clearly, deeper linguistic processings that lead to better document or query enrichment and understanding would certainly allow to leverage the results both from a purely textual viewpoint as well as from a cross-media viewpoint.

Acknowledgments. This work was partly funded by the French National Agency under the *OMNIA* project (OMNIA ANR-07-MDCO-009-01) and under the *Infomagic* project, part of the Pole CAP DIGITAL (IMVN) de Paris

References

1. Clinchant, S., Renders, J.M., Csurka, G.: Trans-media pseudo-relevance feedback methods in multimedia retrieval. In: *Advances in Multilingual and Multimodal Information Retrieval*. LNCS (2007) 569–576
2. Ah-Pine, J., Cifarelli, C., Clinchant, S., Csurka, G., Renders, J.: XRCE's participation to imageCLEF 2008. In: *Working Notes of the 2008 CLEF Workshop*, Aarhus, Denmark (2008)
3. Ponte, J., Croft, W.: A language modelling approach to information retrieval. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Information Retrieval*, Melbourne, Australia, ACM (1998) 275–281
4. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: *Proceedings of IEEE CVPR Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, USA. (2007)
5. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for re-ordering documents and producing summaries. In: *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, Melbourne, Australia, ACM (1998) 335–336
6. Michaud, P., Marcotorchino, F.: Modeles d'optimisation en analyse des données relationnelles. *Math. Sci. Hum.* **67** (1979) 7–38
7. Marcotorchino, J., Michaud, P.: Heuristic approach of the similarity aggregation problem. *Methods of operation research* **43** (1981) 395–404
8. Arni, T., Clough, P., Sanderson, M., Grubinger, M.: Overview of the ImageCLEF-photo 2008 photographic retrieval task. In: *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum. Lecture Notes in Computer Science*, Aarhus, Denmark (September 2008 (printed in 2009))