A Comparative Study of Diversity Methods for Hybrid Text and Image Retrieval Approaches

Sabrina Tollari[1], Philippe Mulhem[2], Marin Ferecatu[3], Hervé Glotin[4],
Marcin Detyniecki[1], Patrick Gallinari[1], Hichem Sahbi[3], and Zhong-Qiu Zhao[4,5]

[1] Université Pierre et Marie Curie - Paris 6, UMR CNRS 7606 LIP6, Paris,
firstname.lastname@lip6.fr
[2] Université Joseph Fourier, UMR CNRS 5217 LIG, Grenoble, first.lastname@imag.fr
[3] TELECOM ParisTech, UMR CNRS 5141 LTCI, Paris,
firstname.lastname@telecom-paristech.fr
[4] Université du Sud Toulon-Var, UMR CNRS 6168 LSIS, Toulon, name@univ-tln.fr
[5] Computer and Information School, Hefei University of Technology, China

Abstract. This article compares eight different diversity methods: 3
based on visual information, 1 based on date information, 3 adapted to
each topic based on location and visual information; finally, for com-
pleteness, 1 based on random permutation. To compare the effectiveness
of these methods, we apply them on 26 runs obtained with varied meth-
ods from different research teams and based on different modalities. We
then discuss the results of the more than 200 obtained runs. The results
show that query-adapted methods are more efficient than non-adapted
method, that visual only runs are more difficult to diversify than text
only and text-image runs, and finally that only few methods maximize
both the precision and the cluster recall at 20 documents.

1 Introduction

Information retrieval is generally based on the computation of similarity with
the query. Thus very similar images appearing in almost identical documents
are retrieved with comparable degrees, producing clusters of alike images alike
in the final ranking. In order to reduce this redundancy, several research teams
[1–3, 5–7] propose to apply, after retrieving the images, a diversity method.

The 2008 ImageCLEFphoto task [1] was focused on diversity. The evaluation
was based on two measures: precision at rank 20 (P20) and instance recall at
rank 20 (also called cluster recall (CR20) or S-recall), which calculates the per-
centage of *different* classes or clusters represented in the top 20 results. The idea
behind these measures was to focus on relevant but at the same time diverse -
in terms of clusters - images. Since it is important to maximize simultaneously
both measures, for the overall ranking, the F1-measure (harmonic mean of P20
and CR20) was used.

The number of parameters in play during a retrieval process makes the study
of the diversification approaches complex. For instance, the diversity (measured

by the cluster recall) depends on the classical recall of the different underlying retrieving methods. To successfully compare the effectiveness of diversification methods, it is important to compare them using different multimodal non-diversified runs. In this paper we propose to study eight diversification methods applied to 26 ranked lists obtained with varied methods from different research teams [2, 3, 5, 6] and based on different modalities (text, image, both...).

In Section 2, we briefly describe the characteristics of the eight diversity methods. In Section 3, we first present the 26 non-diversified runs used, then we discuss the results of the application of the diversity methods on the non-diversified runs. Finally we conclude the paper in Section 4.

2   The Diversity Methods

We study two kinds of diversity methods: (i) those where we apply the same diversity criterion for each topic, and (ii) those where the diversity criterion is adapted to the topic in function of the <cluster> field. For more details, please refer to the specific papers listed below.

2.1   The Non-Adapted Diversity Methods

We study three kinds of non-adapted diversity methods: one based on the <date> field, three based on visual information only and one based on random permutation.

`ClustDMY`[**5**] It uses an approach in which all images taken at the same date (day-month-year) are grouped together. All the images that have the same month and year but do not have a day specified are grouped together. All the images that have the same year, but no month and no day are grouped together.

`AffProp`[**3**] First, a clustering on the top 1000 images is performed using affinity propagation and setting the parameters for 20 clusters. Then, images with the lowest rank are selected in each cluster, the remaining images are put in their original rank order.

`DIVVISU`[**6**] A visual space partition of 256 clusters is obtained by binarization of the 8-bin Hue histogram (Hue from HSV space). Images are reranked in order to have each image of the top 20 belonging to a different cluster.

`VisKmeans`[**5**] A KMeans clustering based on the visual description of the images (4608 dimension histograms). The number of clusters is 500. There are on average 40 images per cluster.

`DIVALEA`[**6**] In order to have a point of comparison, this naive method proposes to randomly permute the first 40 retrieve images.

2.2   The Query-Adapted Diversity Methods

The <cluster> field of ImageCLEFphoto 2008 is related to the location of the pictures in 26 out of 39 topics. For the following methods, the diversity is based

on text field for those 26 topics, and based on visual information for the 13 remaining.

Kmeans[5] Queries having a cluster *city* are diversified using a clustering based on the city name coming from the <LOCATION> field of the image descriptions. Queries having a cluster *country* are diversified using a clustering based on the country name coming from the <LOCATION> of image descriptions. Queries having another cluster name are diversified using the `VisKmeans` diversity method (see before).

MAXMIN[2] For the 26 topics related to location, text clustering is used. Otherwise, the diversity method based on visual descriptors is used. The `MAXMIN` diversity algorithm is based on the maximization of the smallest visual distance of a given document with respect to the so far selected results. Let E be the candidate set (the 40 best elements in our experiments) and consider $\mathcal{C} \subset \mathcal{S}$ as the set of already selected examples, the next document is then chosen as $x = \arg\max_{x_k \in \mathcal{S} \setminus \mathcal{C}} \min_{x_i \in \mathcal{C}} d(x_k, x_i)$ This procedure generates a permutation of the relevant query results such as its prefix corresponds to the most diversified results regarding the distance defined in the description space.

QT[2] Text clustering is the same as the `MAXMIN` method. Visual diversity is inspired both by the Quality Threshold clustering and Voronoi algorithms. Let $s = N - n_{\mathcal{C}}$ be the cluster size, where $N, n_{\mathcal{C}}$ are respectively the number of images and the expected number of clusters. The algorithm iteratively updates a list of Voronoi cell prototypes by minimizing the following criterion $y_l = \arg\min_{x_i} \mathcal{R}(KNN_s(x_i; \mathcal{S}_t))$ where $KNN_s(x_i; \mathcal{S}_t$ denotes the s nearest neighbors of x_i in $\mathcal{S}_t$ ($\mathcal{S}_1 = \{x_1, \cdots, x_N\}$) and $\mathcal{R}$ the radius of the smallest ball enclosing $KNN_s(x_i; \mathcal{S}_t$. The new generated Voronoi cells (denoted $\mathcal{C}_l$) are removed and the process is iterated on the remaining data $\mathcal{S}_{l+1} - \mathcal{S}_l \setminus \mathcal{C}_l$. The final result is a partition of Voronoi cells and their prototypes which corresponds to the most diverse results.

The `AffProp`, `ClustDMY`, `DIVVISU`, `VisKmeans`, `Kmeans` and `QT` methods are based on clustering contrary to `DIVALEA` and `MAXMIN` which are based on permutation.

3 Experiments

3.1 The Non-Diversified Runs

In order to make a true comparison of the diversity methods, we propose to apply these methods on 26 different non-diversified runs[6].

Automatic Text Only Runs There are 3 text only runs, each from a different team. The first is based on language modelling [5], the second on tf-idf [6], and the last one uses a cosine measure between topics and matching candidate documents [2].

---

[6] The 26 non-diversified runs and the 208 runs generated by the proposed diversity methods are available at `http://aveir.lip6.fr/diversity`.

Table 1. Comparison of the information used for each diversity method

	Diversity Method	Text	Visual	Other
1	ClustDMY	\<DATE\>	-	-
2	AffProp	-	20 visual clusters built by affinity propagation	-
3	DIVVISU	-	256 clusters (Hue)	-
4	VisKmeans	-	500 clusters (RGB)	-
5	DIVALEA	-	-	random
6	Kmeans	\<LOCATION\>	500 clusters (RGB)	-
7	MAXMIN	\<LOCATION\>	maximising the visual distances	-
8	QT	\<LOCATION\>	20 visual clusters built iteratively	-

Automatic Visual Only Runs There are also 3 visual runs from three different teams. The first is based on grid segmentation and Jensen-Shannon divergence [5], the second on entropic visual features and a 2-class SVM learning machine trained with the Gaussian kernel [3], and the last one on global color, texture and shape visual descriptors and a 2-class SVM learning machine trained with the Laplacian kernel [2].

Automatic Text-Image Runs There are 9 automatic text-image runs. Four runs are from each of the different teams, and the last five ones are the AVEIR fusion runs (see also [7] for more details). The four text-image individual runs are:

LIG: The first run is based on the linear combination of the scores provided by a language model using Dirichlet smoothing on the text and by a Jeffrey-Divergence correspondence on the images [5].

LIP6: In the second run, text processing is based on standard TF-IDF with cosine similarity. Forest of Fuzzy Decision Trees (FFDT) trained on VCDT ImageCLEF task 2008 are used for a visual concept filtering of the textual results. The matching of the concepts and the topics text used WordNet [6].

LSIS: In the third run, the visual features are entropic features. Lots of SVMs are trained and generated with different parameters using the sample images provided. Then the first 20 images of the LIG run are used as the positive samples for each topic, and the others as the negative samples to construct the validation set for selecting the best one among the generated SVMs [3].

PTECH: The last run uses a combination of text and image descriptors. For a given topic, a separate query is performed for each modality (text and image). The results are merged by a minimum rank criterion: each image keeps the best rank [2].

The five AVEIR runs [7] correspond to different fusion strategies applied to the four text-image individual runs described before. Those runs were well ranked in the ImageCLEFphoto 2008 competition[7].

MIN: for each image, the fusion-rank corresponds to the minimum rank observed on each of the 4 team's runs. This strategy corresponds to creating a rank by alternatively choosing an image from each of the teams' runs. The first image of the fusion rank corresponds to the first image of the first team; the second image corresponds to the first image of the second team; the fifth corresponds to the second image of the first team, and so on.

MEAN: for each image, the fusion-rank corresponds to the average rank observed on each of the 4 team's runs. This strategy corresponds to a compromise taking into account all the systems. Images not present in one of the ranked lists are considered as having rank 1001.

MEAN1on4: here only images that are ranked by at least 1 teams were considered. The fusion-rank correspond to the average of the available ranks.

MEAN2on4: same as MEAN1on4, but only images that are ranked by at least two teams were considered. The idea behind this strategy is to avoid fusion of images returned only by one team.

MEAN3on4: same as MEAN1on4, but only images that are ranked by at least three teams were considered. The idea behind this strategy is to avoid fusion of images returned only by two teams.

Manual Text-Image Run There is only one manual run proposed in [2]. For this run, text descriptions are built using the vector space model, while topics are represented as boolean queries built manually. The visual and text results are combined by intersecting the underlying results, so this will guarantee that the output is consistent with respect to both modalities.

Ideal Runs To measure to what extent our diversity methods are efficient independently of the runs there are applied on, we also propose to build ideal runs. Using the ground truth, we build the set of relevant images for each topic, then we randomly permute those images to obtain the ideal runs. We reiterate the process 10 times in order to obtain 10 ideal runs. Obviously, the precision at 20 documents will be closer to 1, because all the retrieved documents are relevant, but the cluster recall at 20 documents will not be optimal.

3.2 Results

When averaging on the 10 ideal runs (see Figure 1(a)), `Kmeans`, `MAXMIN` and `QT` give the best results for the CR20 (above 0.84) and for the F1-measure (above 0.90). For ideal runs, we consider that checking the precision at 20 documents is not relevant (P20=0.993 ± 0.001), because reordering only relevant documents lead to the same value. It is worth noting that, on average on the 10 ideal runs,
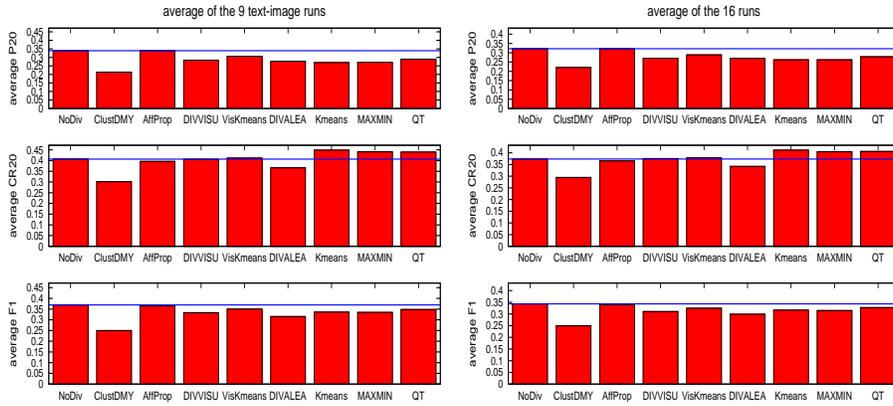
---

[7] Particularly, the MEAN run is ranked 18 under 1042 submitted runs (P20=0.43, CR20=0.46 and F1-measure=0.45) (see [7] for more details)
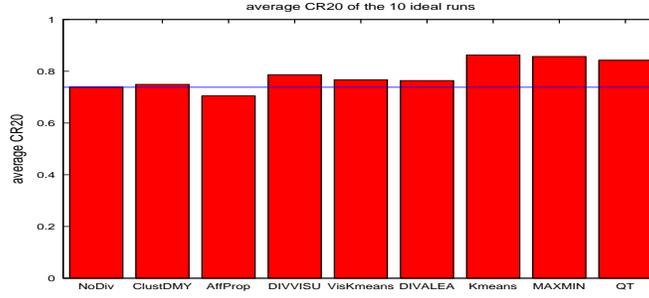
(a) Average CR20 on the 10 ideal runs (P20=0.993 ± 0.001)
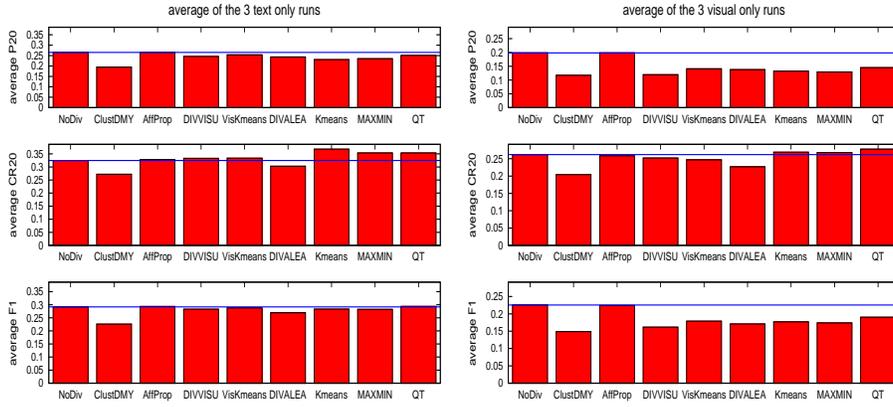
(b) Average P20, average CR20 and average F1 measure of the 3 text only runs

(c) Average P20, average CR20 and average F1 measure of the 3 visual only runs

(d) Average P20, average CR20 and average F1 measure of 9 text-image runs

(e) Average P20, average CR20 and average F1 measure of the 16 runs

Fig. 1. Comparison of the diversity methods in function of the modality of the runs. First bar (and line) corresponds to the value without diversity

the random reranking `DIVALEA` increases the CR20 results, and such reranking also outperforms for the CR20 value the `ClustDMY` and `AffProp`, for which we do not have yet a clear explanation.

When considering only the text only runs (see Figure 1(b)), `AffProp`, `VisKmeans` and `QT` achieves a F1-measure value as high as the non diversified runs. An important point to notice here is that all the diversification schemes (except `ClustDMY` and `DIVALEA`) outperform the non diversified result of 0.325 for the CR20, with values higher than 0.35 for `Kmeans`, `QT` and `MAXMIN`.

When considering the three visual only runs (see Figure 1(c)), the P20 scores strongly decrease except for `AffProp`where the P20 score is similar with the non-diversified run. Even if the the P20 scores strongly decrease, the CR20 scores for `Kmeans`, `MAXMIN` and `QT` are above the CR20 of the non-diversified run. But finally the F1-measures are very low. This leads to the fact that diversifying visual only runs is much more difficult to achieve.

For the automatic text-image runs (see Figure 1(d)), the non diversified run outperforms the diversified runs for the F1-measure. One again, the `Kmeans`, `QT` and `MAXMIN` diversifications give better results that the non diversified runs for the CR20 value. The same conclusion are inferred from the 5 AVEIR runs, from the 4 individual text-image runs and from the manual run.

For all the runs (image only, text only and image-text) without the ideal runs (see Figure 1(e)), the `Kmeans`, `MAXMIN` and `QT` diversity methods outperforms the CR20 of the non diversified run; `AffProp`, `DIVVISU` and `VisKmeans` give similar CR20 than the non diversified run, we conclude from these results that diversity based only on visual information is not effective; `DIVALEA` decreases the CR20; clustering on day/month/year `ClustDMY` gives also bad results, this may be due to the fact that all the images do not have the full date given. Finally, if we compare the F1-measure of the 16 runs (see Figure 1(e)), we noted that the F1-measure values of the diversified runs are always below the F1-measure value of the non-diversified run. We conclude that it is very hard to have a high CR20 score and at the same time a high P20 score. We only achieve this goal with the query-adapted diversity methods and the ideal runs (see Figure 1(a)).

4   Conclusion

In this article, we compare eight diversity methods applied on 26 varied runs from different teams. This study shows that for all the diversity methods, the precision at 20 always decreases compared to the run without diversification (except - of course - for ideal runs).

The results first show that query-adapted methods (`Kmeans`, `MAXMIN` and `QT`) are more efficient than non-adapted methods. But, in the case of real web search engines, it is difficult for a user to choose the right diversity criterion to apply to a query, so even if non-adapted methods are less efficient they must also be considered. The diversity based on the date `ClustDMY` appears to be non efficient and the visual-only diversity methods (`AffProp`, `VisKmeans` and `DIVVISU`) gives similar cluster recall than the non diversified results. Diversifying a ranked list

of image results when no diversity cluster is given is always an open question. Second, because visual only runs are more difficult to diversify than text only and text-image runs, diversifying content based image retrieval results are more difficult than diversifying text-based image retrieval. Finally, we unfortunately notice that, in the case of 16 non ideal runs, none of the eight diversity methods obtains a better F1-measure value than the non-diversified run. It is only in the case of the 10 ideal runs that some of the diversity methods gives a better F1-measure value. In future work, we will also compare the effect of diversity based on visual concept [4].

Acknowledgment

References

1. T. Arni, P. Clough, M. Sanderson, and M. Grubinger. 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, 2008.
2. M. Ferecatu and H. Sahbi. TELECOM ParisTech at ImageClefphoto 2008: Bimodal text and image retrieval with diversity enhancement. In *Working Notes for the CLEF 2008 workshop*, 2008.
3. H. Glotin and Z. Zhao. Visual-only affinity propagation promoting diversity for CLEF2008 photo retrieval task. In *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, LNCS, 2008.
4. M. Inoue and P. Grover. Effects of visual concept-based post-retrieval clustering in imageclefphoto 2008. In *Working Notes for the CLEF 2008 workshop*, 2008.
5. L. Maisonnasse, P. Mulhem, Eric Gaussier, and J.-P. Chevallet. Lig at ImageCLEF. In *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, Lecture Notes in Computer Science, 2008.
6. S. Tollari, M. Detyniecki, A. Fakeri-Tabrizi, Massih-Reza Amini, and P. Gallinari. Using visual concepts and fast visual diversity to improve image retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access – 9th Workshop of the Cross-Language Evaluation Forum*, LNCS, 2008.
7. S. Tollari, M. Detyniecki, M. Ferecatu, H. Glotin, P. Mulhem, M.-R. Amini, A. Fakeri-Tabrizi, P. Gallinari, H. Sahbi, and Z.-Q. Zhao. Consortium AVEIR at ImageCLEFphoto 2008: on the fusion of runs. In *Working Notes for the CLEF 2008 workshop*, 2008.