UPMC at MediaEval 2013: Relevance by Text and Diversity by Visual Clustering

Christian Kuoman
XILOPIX
88000 Epinal, France
christian@xilopix.net

Sabrina Tollari
UPMC Univ Paris 06 / LIP6
75005 Paris, France
Sabrina.Tollari@lip6.fr

Marcin Detyniecki
UPMC Univ Paris 06 / LIP6
75005 Paris, France
Marcin.Detyniecki@lip6.fr

ABSTRACT

In the diversity task, our strategy was to, first, try to improve relevance, and then to cluster similar images to improve diversity. We propose a four step framework, based on AHC clustering and different reranking strategies. A large number of tests on devset showed that most of the best strategies include text based reranking for pertinence, and visual clustering for diversity - even compared to location based descriptors. Results on expert and crowd-sourcing testset grounds truths seem to confirm these observations.

1. INTRODUCTION

In the Retrieving Diverse Social Images Task of MediaEval 2013 challenge [1], participants were provided with a ranked list (we called "baseline") of at most 150 photos of a location (the query) from Flickr.com. For each query, our strategy to induce diversity while keeping the relevance is based on four steps. Step 1: Rerank the baseline to improve relevance. Step 2: Cluster the results using an Agglomerative Hierarchical Clustering (AHC). Step 3: Sort the clusters based on a cluster priority criteria; and then sort the images in each cluster. Step 4: Finally rerank the results alternating images from different clusters.

It is important to notice that the AHC does not take the image rank into account, but when we sort the clusters and the images in the cluster (Step 3) the rank obtained in Step 1 is crucial information that we exploit. In fact, it is the only way to guarantee global relevance with respect to the query.

2. SIMILARITIES AND DISTANCES

To rerank the baseline list according to the similarity to the query (Step 1) and to cluster images (Step 2), we need to compare the images. We tested on the devset several similarities and distances, for different types of descriptors: visual, textual, GPS and a geographic tree thesaurus. For all visual descriptors provided by the organizers [1] (CN3x3, LBP3x3, CSD, HOG...), we use the Euclidean distance. For textual descriptors, we use Dirichlet Prior Smoothing [3] for the probabilistic model; the cosinus for TF-IDF weighting; the formula mentioned in [4] for Social TF-IDF weighting.

To estimate the distance between two GPS coordinates, we compute the classical great-circle distance using the Haversine formula. For the keywordsGPS subset, all the retrieved images have GPS coordinates; but for the keywords subset, approximately 60% of the images do not have any coordinates. For these images, we choose to attribute them the GPS coordinates of the nearest image, among the images of the same query, according to the visual distance. Moreover, if the smallest visual distance is greater than a threshold, the system associates the (0,0) GPS coordinates to the image in order to avoid some noisy results.

To better exploit geographical granularity between images, we use the "thesaurus" developed by the commercial search engine Xilopix (see [2] for details). The "travel" domain of this thesaurus is organized into a tree of concepts: continents, countries, regions, departments and locations. For each concept, the thesaurus provides its name and its GPS coordinates. For images with GPS, the system calculates the great-circle distance between the GPS coordinates of the image and the GPS coordinates of each concept in the thesaurus, and finally selects the closest concept and its parent nodes (method called *tree*). For images without GPS, the system matches the terms of the image and the terms of each thesaurus node using TF-IDF weighting (method called *tree-tfidf*) or probabilistic models (method called *tree-proba*) and finally selects the closest concept and its parent nodes. To estimate the similarity between two concepts in the thesaurus, we use the Wu-Palmer's similarity [5] that quantifies the similarity between two concepts of a same tree.

3. CLUSTERING BY AHC

The Agglomerative Hierarchical Clustering (AHC) is a clustering method that provides a hierarchy of clusters of images. Applying the AHC to the query results provides a dendrogram. In order to obtain groups of similar images, we choose to cut the dendrogram to obtain a fixed number of unordered clusters (method called *FixedN* where *N* is the number of clusters to obtain) (see [2] for more details).

The way most diversity methods work implies what we call "rank priority" (*rank*): we first choose the cluster containing the image of rank 1 in Step 1, then we choose a *different* cluster containing the next possible lowest rank. Other ways to prioritize the clusters may be interesting, we propose to consider the number of images contained in each cluster. We sort the clusters in decreasing order from the cluster with the largest numbers of images to the cluster with the less images (*dec* priority). After sorting the clusters, we sort the images in each cluster according to their rank in Step 1.

4. EXPERIMENTS AND RESULTS

On devset, we tested our model for all descriptors and for

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

Table 1: Submitted runs: parameters (top), results on devset (middle) and results on testset (bottom). Between brackets, gain in percentage compared with run1_visual

	keywordsGPS			keywords		
	SUBMITTED RUNS PARAMETERS					
	Rerank	AHC	Priority	Rerank	AHC	Priority
run1_visual	baseline	CN3x3	dec, fixed35	LBP3x3	CSD	dec, fixed20
run2_text	tfidf(tt)	proba(tt)	rank,fixed35	tfidf(ttd)	social-tfidf(ttd)	rank,fixed25
run3_textvisual	tfidf(tt)	CSD	dec, fixed20	tfidf(ttd)	CSD	rank,fixed25
run5_allallowed	tree	CSD	dec, fixed30	tfidf(ttd)	tree-proba(ttd)	rank,fixed15
	RESULTS ON DEVSET					
number of queries	25			25		
	P@10	CR@10	F1@10	P@10	CR@10	F1@10
baseline	0.860 (-)	0.412 (-)	0.544 (-)	0.688 (-)	0.464 (-)	0.529 (-)
run1_visual	0.868(ref)	0.498(ref)	0.623(ref)	0.696(ref)	0.543(ref)	0.575(ref)
run2_text	**0.928(+7)**	0.493(-1)	**0.636(+2)**	0.788(+13)	**0.586(+8)**	0.629(+9)
run3_textvisual	0.844(-3)	**0.509(+2)**	0.627(+1)	**0.812(+17)**	**0.584(+8)**	**0.635(+10)**
run5_allallowed	0.808(-7)	0.483(-3)	0.592(-5)	0.760(+9)	0.560(+3)	0.594(+3)
	RESULTS ON TESTSET					
number of queries	210			132		
	P@10	CR@10	F1@10	P@10	CR@10	F1@10
run1_visual	0.774(ref)	0.370(ref)	0.489(ref)	0.630(ref)	0.400(ref)	0.468(ref)
run2_text	**0.844(+9)**	0.404(+9)	0.531(+9)	**0.746(+18)**	**0.412(+3)**	**0.507(+8)**
run3_textvisual	0.823(+6)	**0.426(+15)**	**0.547(+12)**	0.718(+14)	**0.417(+4)**	**0.503(+8)**
run5_allallowed	0.766(-1)	0.378(+2)	0.496(+1)	0.705(+12)	0.388(-3)	0.475(+2)

Table 2: Scores obtained for 3 crowd-sourcing grounds truths (GT1, GT2, GT3) and for the expert ground truth (GT0). All the results are in average for a subset of 49 queries of testset. *nb* is the number of queries among the 49 queries which have a CR@10=1. Between brackets, gain in % compared with run1_visual

	GT1,2,3	GT1		GT2		GT3		GT0		
	P@10	CR@10	*nb*	CR@10	*nb*	CR@10	*nb*	P@10	CR@10	*nb*
run1_visual	0.694(ref)	0.786(ref)	27	0.754(ref)	10	0.645(ref)	14	0.806(ref)	0.367(ref)	0
run2_text	**0.757**(+9)	0.836(+6)	31	0.756(+0)	16	0.645(-0)	14	**0.851**(+6)	0.408(+11)	1
run3_textvisual	0.749(+8)	**0.886**(+13)	33	**0.792**(+5)	21	**0.687**(+6)	16	0.841(+4)	**0.415**(+13)	0
run5_allallowed	0.708(+2)	0.828(+5)	29	0.768(+2)	20	0.643(-0)	15	0.794(-2)	0.377(+3)	0

most of the parameters. For textual models, on keywords subset, we choose to use the title, tags and descriptions (ttd) fields, while on keywordsGPS, we choose to use only the title and tags (tt) fields. Among our large number of tests, Table 3 shows an example of comparison of AHC results on devset keywordsGPS for tree, GPS and visual (CSD) descriptors using the same parameters and the same Step 1 reranking approach (i.e. tfidf(tt)). Best diversity results are obtained with visual descriptors compared to tree and GPS.

According to the results on devset, we choose the methods and the parameters for each subset. Table 1 summarizes the parameters and the scores obtained on devset and testset according of the expert ground truth, while Table 2 compares the results on the crowd-sourcing grounds truths.

Table 3: Comparison on devset keywordsGPS

Rerank	AHC	Priority	P@10	CR@10
baseline	-	-	0.860(ref)	0.412(ref)
tfidf(tt)	-	-	0.896(+4)	0.429(+4)
tfidf(tt)	tree	dec,fixed30	0.840(-2)	0.438(+6)
tfidf(tt)	GPS	dec,fixed30	0.864(+0)	0.443(+8)
tfidf(tt)	CSD	dec,fixed30	0.864(+0)	**0.485**(+18)

5. CONCLUSION AND DISCUSSION

Results on expert and crowd-sourcing grounds truths suggest that an interesting and robust strategy to improve diversity - in the sense of this challenge - is to increase the relevance using the text, and then to exploit visual clustering to diversify the results. Preliminary tests on devset showed that the exploitation of these descriptors outperforms, in terms of diversity, the use of location descriptors (GPS or tree). This is an unexpected results taking into account that queries were formulated around the notion of location.

6. REFERENCES

[1] B. Ionescu, M. Menéndez, H. Müller, and A. Popescu. Retrieving diverse social images at MediaEval 2013: Objectives, dataset and evaluation. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[2] C. Kuoman, S. Tollari, and M. Detyniecki. Using tree of concepts and hierarchical reordering for diversity in image retrieval. In *CBMI*, pages 251–256, 2013.

[3] T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[4] A. Popescu and G. Grefenstette. Social media driven image retrieval. In *ACM ICMR*, 2011.

[5] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *Ass. for Computational Linguistics*, 1994.