LAPI @ 2015 Retrieving Diverse Social Images Task: A Pseudo-Relevance Feedback Diversification Perspective

Bogdan Boteanu[1]*, Ionuţ Mironică[1]†, Bogdan Ionescu[1]‡
[1]LAPI, University "Politehnica" of Bucharest, Romania
{bboteanu,imironica,bionescu}@alpha.imag.pub.ro

ABSTRACT

In this paper we present the results achieved during the 2015 MediaEval Retrieving Diverse Social Images Task, using an approach based on pseudo-relevance feedback, in which human feedback is replaced by an automatic selection of images. The proposed approach is designed to have in priority the diversification of the results, in contrast to most of the existing techniques that address only the relevance. Diversification is achieved by exploiting a hierarchical clustering scheme followed by a diversification strategy. Methods are tested on the benchmarking data and results are analyzed. Insights for future work conclude the paper.

1. INTRODUCTION

An efficient information retrieval system should be able to provide search results which are in the same time *relevant* for the query and cover different aspects of it, i.e., *diverse*. The 2015 Retrieving Diverse Social Images Task [1] addresses this issue in the context of a tourism real-world usage scenario. Given a ranked list of location photos retrieved from Flickr[1], participating systems are expected to refine the results by providing up to 50 images that are in the same time relevant and provide a diversified summary of the location. These results will help potential tourists in selecting their visiting locations. The refinement and diversification process is based on the social metadata associated with the images and/or on the visual characteristics. A complete overview of the task is presented in [1].

Despite the current advances of machine intelligence techniques used in the area of information retrieval and multimedia, in search for achieving high performance and adapting to user needs, more and more research is turning now towards the concept of "*human in the loop*" [2]. The idea is to bring the human expertise in the processing chain, thus combining the accuracy of human judgements with the computational power of machines.

In this work we propose a novel perspective that exploits the concept of pseudo-relevance feedback (RF). RF techniques attempt to introduce the user in the loop by harvesting feedback about the relevance of the search results. This information is used as ground truth for re-computing a better representation of the data needed.

*This work has been funded by the Ministry of European Funds through the Financial Agreement POSDRU 187/1.5/S/155420.
†The work was funded by the ESF POSDRU/159/1.5/S/132395 InnoRESEARCH programme.
‡This work is supported by the European Science Foundation, activity on "Evaluating Information Access Systems".
[1]http://flickr.com/.

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

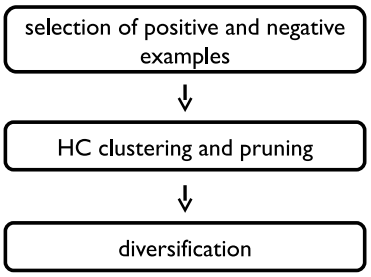

Figure 1: General scheme of the proposed approach

Relevance feedback proved efficient in improving the precision of the results [3], but its potential was not fully exploited to diversification. The main contribution of our approach is in proposing a pseudo-relevance feedback technique which substitutes the user needed in traditional RF and in proposing several diversity-adapted relevance feedback schemes.

2. PROPOSED APPROACH

In traditional RF Techniques, recording actual user feedback is inefficient in terms of time and human resources. The proposed approach, denoted in the following *HC-RF*, attempts to replace user input with machine generated ground truth. It exploits the concept of pseudo-relevance feedback. The concept is based on the assumption that top k ranked documents are relevant and the feedback is learned as in traditional RF under this assumption [6]. A general diagram of the approach is depicted in Figure 1.

The algorithm is as follows. Firstly, we remove non-relevant images using three filters. The first one is the Viola-Jones [4] *face detector*, which filters out images with persons as the main subject. Second one is an image blur detector based on the aggregation of 10 state-of-the-art blur indicators as implemented by Said Pertuz[2]. The last one is a GPS distance-based filter, which rejects the images that are positioned too far away from the query location, and therefore which cannot be relevant shots for that location.

In the next step we propose a pseudo-relevance feedback scheme based on the selection of the images assessed in an automated manner. We consider that most of the first returned results are relevant (i.e., positive examples). For instance, on *devset* [1], in average, 40 out of 50 returned images are relevant which support our assumption. In contrast, the very last of the results are more likely non-relevant and considered accordingly (i.e., negative examples).

[2]http://www.mathworks.com/matlabcentral/fileexchange/27314-focus-measure/content/fmeasure/fmeasure.m

Table 1: Best pseudo-relevance feedback results for each modality or combination of modalities on devset (best results are depicted in bold).

metric/ method	HC-RF visual	HC-RF text	HC-RF vis-text	HC-RF cred.	HC-RF CNN	Flickr init. res.
P@20	0.8199	**0.8346**	0.8281	0.7281	0.7546	*0.8118*
CR@20	0.4423	**0.4588**	0.4484	0.4415	0.4234	*0.3432*
$F1$@20	0.5655	**0.5839**	0.5735	0.5426	0.5356	*0.4713*

The positive and negative examples are feed to an Hierarchical Clustering[3] scheme which yields a dendrogram of classes. For a certain cutting point (i.e., number of classes), a class is declared non-relevant if contains only negative examples or the number of negative examples is higher than the positive ones. The final step is the actual diversification scheme. We select from each of the relevant classes one image which has the highest rank according to the initial ranking of the system. Then we proceed by selecting the second image in the same manner and the process is repeated until a maximum number of images is reached. The resulting images represent the output of the proposed system.

3. EXPERIMENTAL RESULTS

This section presents the experimental results achieved on *devset* which consists of 153 queries and 45,375 images and *testset*, respectively, which consists in 139 queries (69 one-concept - 70 multi-concept) and 41,394 images. For *devset*, we first optimized the parameters of the filters in order to obtain best precision. Based on this configuration we then applied the proposed approach. Ground truth was also provided with the data for this set for preliminary validation of the approaches. The final benchmarking is conducted however on *testset*.

In our approaches, images are represented with the content descriptors that were provided with the task data, i.e., visual (e.g., color, feature descriptors), text (e.g., term frequency - inverse document frequency representations of metadata) and user annotation credibility (e.g., face proportions, upload frequency) information. Detailed information about provided content descriptors is available in [1]. Performance is assessed with Precision at X images (P@X), Cluster Recall at X (CR@X) and F1-measure at X (F1@X).

3.1 Results on devset

Several tests were performed with different descriptor combinations and various cutoff points. Descriptors are combined with an early fusion approach. We varied the number of initial images considered as positive examples, from 80 to 160 with a step of 10 images, the number of last images considered as negative examples, from 0 to 21 with a step of 3, and the inconsistency coefficient threshold for which HC naturally divides the data into well-separated clusters, from 0.1 to 0.95 with a step of 0.05. We select the combinations yielding the highest $F1$@20, which is the official metric.

While experimenting, we observed that, by increasing the number of analyzed images, precision tends to slightly decrease as the probability of obtaining un-relevant images increases; in the same time, diversity increases as having more images is more likely to get more diverse representations. For brevity reasons, in the following we focus on presenting only the results at a cutoff of 20 images which is the official cutoff point. These results are presented in Table 1. To serve as baseline for the evaluation, we present also the Flickr initial retrieval results. From the modality point of view, text descriptor (TF) lead to the highest results ($F1$@20=0.5839) followed closely by the combination of all visual and all text descriptors ($F1$@20=0.5735) and then visual (LBP) ($F1$@20=0.5655), all credibility information ($F1$@20=0.5426) and all convolutional neural network (CNN) based descriptors ($F1$@20=0.5356).

[3] http://www.mathworks.com/help/stats/hierarchical-clustering.html

Table 2: Results for the official runs on testset (best results are depicted in bold).

set	metric	*Run1*	*Run2*	*Run3*	*Run4*	*Run5*
Overall	P@20	0.7241	0.709	**0.7306**	0.7126	0.7227
	CR@20	0.4156	0.4306	0.4062	**0.449**	0.3999
	$F1$@20	0.5164	0.5231	0.5056	**0.5336**	0.4994
One-topic	P@20	0.7319	0.7391	0.7341	**0.7442**	0.7123
	CR@20	0.4153	**0.4392**	0.4211	0.4294	0.3934
	$F1$@20	0.5222	**0.5402**	0.5219	0.5308	0.4958
Multi-topics	P@20	0.7164	0.6793	0.7271	0.6814	**0.7329**
	CR@20	0.416	0.4222	0.3915	**0.4684**	0.4063
	$F1$@20	0.5108	0.5063	0.4895	**0.5364**	0.503

3.2 Official results on testset

Following the previous experiments, the final runs were determined for best modality/parameter combinations obtained on *devset* (see Table 1). We submitted five runs, computed as following: *Run1* - automated using visual information only: HC-RF visual LBP, *Run2* - automated using text information only: HC-RF text TF, *Run3* - automated using visual-text information: HC-RF all visual-all text, *Run4* - automated using credibility information only: HC-RF all cred., and *Run5* - everything allowed: HC-RF all CNN. Results are presented in Table 2.

What is interesting to observe is the fact that the highest precision is achieved on *one-topic* set, using credibility information, (*Run4* - $P@20 = 0.7442$), whereas maximum diversification is achieved on *multi-topics* set, using the same type of information (*Run4* - $CR@20 = 0.4684$). Another interesting observation is that credibility information was useful in the context of overall diversification. Credibility information gives an automatic estimation of the quality of tag-image content relationships, telling which users are most likely to share relevant images in Flickr. Best diversification is achieved, $CR@20 = 0.4684$, due to the high probability that different relevant images belong to different users with a good credibility score. In terms of $F1$ metric score, the use of credibility information, *Run4* - $F1@20 = 0.5336$, allows for better performance over text descriptor (TF) by almost 1% and by 1.7% over visual descriptor (LBP).

4. CONCLUSIONS

We approached the image search result diversification issue from the perspective of relevance feedback techniques, when user feedback is substituted with an automatic pseudo-feedback approach. Results show that in general, the automatic techniques improve the precision and diversification, which proves the real potential of relevance feedback to the diversification. Future developments will mainly address a more efficient exploitation of different modalities (visual-text-credibility), e.g., via late fusion techniques, as well as exploitation of adaptive face-detectors that are able to filter out only a certain category of images, e.g., with people in focus, and pass other categories of images, e.g., with crowds that are naturally present at a target location.

5. REFERENCES

[1] B. Ionescu, A.L. Gînscă, B. Boteanu, A. Popescu, M. Lupu, H. Müller, "*Retrieving Diverse Social Images at MediaEval 2015: Challenge, Dataset and Evaluation*", MediaEval 2015 Workshop, September 14-15, Wurzen, Germany, 2015.

[2] B. Emond, "*Multimedia and Human-in-the-loop: Interaction as Content Enrichment*", ACM Int. Workshop on Human-Centered Multimedia, pp. 77-84, 2007.

[3] J. Li, N.M. Allinson, "*Relevance Feedback in Content-Based Image Retrieval: A Survey*", Handbook on Neural Information Processing, 49, pp. 433-469, Springer 2013.

[4] P. Viola, M. J. Jones, "Robust Real-Time Face Detection," in International Journal of Computer Vision, 57(2), pp. 137–154, 2004.

[5] B. Boteanu, I. Mironică, B. Ionescu, "*A Relevance Feedback Perspective to Image Search Result Diversification*", IEEE ICCP, September 4-6, Cluj-Napoca, Romania, 2014.

[6] B. Boteanu, I. Mironică, B. Ionescu, "*Hierarchical Clustering Pseudo-Relevance Feedback for Social Image Search Result Diversification*", IEEE CBMI, June 10-12, Prague, Czech Republic, 2015.